UNIVERSIDAD DE CANTABRIA

Facultad de Ciencias

INFERENCIA DE LOS GENES DEL ANCESTRO DE LOS AMNIOTAS Y SU RELACIÓN CON EL ORIGEN DEL HUEVO

(Inference of the genes in the ancestor of amniotes and the genomic basis for the origin of the egg)

Trabajo de Fin de Máster
para acceder al

MÁSTER EN CIENCIA DE DATOS

Autor: María Lavín Cabanas

Director: Iker Irisarri Aedo

Septiembre – 2019

ABSTRACT

Amniotes are the first fully terrestrial vertebrate animals with several evolutionary innovations in their common ancestor that allowed them to become fully independent of the aquatic environment, including a more complex egg with shell and additional structures. During evolution, organisms' form and functions evolve, and so do their genomes, which are the ultimately responsible for the observed changes. In fact, genomes are evolutionarily labile and experience changes in gene content and structure. This project aims to investigate the genomic basis for the origin of amniotes and their evolutionary innovations. From a bioinformatic point of view, this work involves: (i). choosing the highest quality vertebrate genomes to use in our analysis; (ii). estimating the genes that originated in the common ancestor of reptiles, birds and mammals by searching for sequence similarity and clustering of homologous genes; (iii). functionally characterizing the novel genes that originated in the common ancestor of amniotes and identifying any relationship with the origin of the amniote egg.

Keywords: Amniote, Comparative genomics, Egg, Gene gain and loss, Genomes, Evolution, Homology, Gene ontology.

RESUMEN

Los amniotas son los primeros animales vertebrados completamente terrestres con varias innovaciones evolutivas en su ancestro común que les permitió independizarse totalmente del entorno acuático, entre ellos un huevo más complejo con cáscara y estructuras adicionales. Durante la evolución, el aspecto y la función de los organismos evolucionan, al igual que sus genomas, que son los responsables de las transformaciones observadas. De hecho, los genomas están constantemente sometidos a cambios debido a la evolución y experimentan modificaciones en el contenido y la estructura de los genes. El objetivo principal de este proyecto es investigar la base genómica del origen de los amniotas y sus innovaciones evolutivas. Desde un punto de vista bioinformático, este trabajo implica: (i). seleccionar los genomas de vertebrados de mayor calidad para usarlos en nuestro análisis; (ii). estimar los genes nuevos que se originaron en el ancestro común de reptiles, aves y mamíferos mediante la búsqueda de similitud de secuencias y agrupamiento de genes homólogos; (iii). caracterizar funcionalmente los genes inferidos como nuevos en el ancestro de los amniotas e identificar una posible relación con el origen del huevo amniota.

Palabras clave: Amniota, Genómica comparativa, Huevo, Ganancia y pérdida de genes, Genomas, Evolución, Homología, Ontología de genes.

Great things in business are never done by one person. They're done by a team of people.

— Steve Jobs

ACKNOWLEDGMENTS

A todos aquellos que en algún momento de mi vida han formado parte de mi proceso educativo quiero agradecerles tanto las buenas como las malas lecciones.

Quiero dar las gracias, especialmente a Iker Irisarri, por su disponibilidad y ayuda en toda ocasión a pesar de mis limitados conocimientos en el campo de la biología.

A Aida Palacio y Jesús E. Marco, por proporcionarme acceso a la Supercomputadora Altamira en el Instituto de Física de Cantabria (IFCA-CSIC), miembro de la Red Española de Supercomputación, además de ayudar a solucionar cualquier contratiempo en el menor tiempo posible.

Agradecer a Jordi Paps su ayuda con este pipline PAPS en su publicación previa.

Finalmente, a mis padres y a todos mis amigos que incluso estando lejos nunca han dejado de apoyarme.

CONTENTS

1 INTRODUCTION

The goal of the present opening section is to introduce the reader to the biological questions and the principles of the main bioinformatics tools used in this thesis.

1.1 THE AMNIOTE EGG

Amniotes fit together in a clade which includes nearly all the vertebrates on land these days, i.e. reptiles, birds and mammals. The common ancestor of all these groups is hypothesized to resemble the earliest amniotes. In comparison with amphibians, amniotes are fully terrestrial vertebrates, i.e. they are able to complete their life cycle independently of water bodies. This transition required many adaptations, one of the most remarkable ones being the evolution of a more complex egg structure with shell[1].

The egg is such an important structure that it was one of the main characters used by Haeckel to separate amniotes from amphibians in his taxonomy of the vertebrates[2].

Land vertebrates (Tetrapoda) appeared in the Carboniferous, ca. 350 million years ago (Irisarri et al. 2017[3]). The fossil record shows the appearance of fully formed amphibians with well-developed limbs and other features indicating that they were terrestrial as adults more than 300 million years ago[1]. However, amphibians were, and still are, necessarily dependent on water to complete their life cycles.

Amphibians typically lay eggs on water and are aquatic for the first period of their lives, until metamorphosed[1]. A typical life cycle of the amphibians is presented in Figure 1.1.

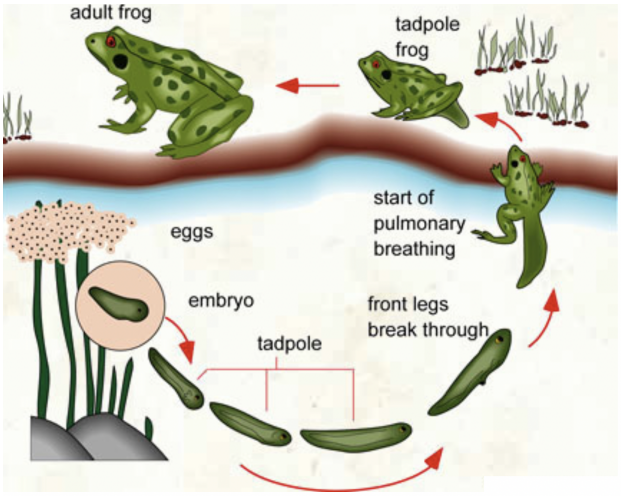

Figure 1.1: Diagram of a life cycle of a frog as an example of amphibian life cycle. Both phases in water and on land are presented[4].

Amphibian eggs do not have shells that protect them from drought but are covered in a jelly-like substance that helps them keep the eggs moist and offers some protection from predators, but need to be laid on water or a wet environment. In addition, amphibian eggs generally contain only a modest amount of yolk and do not develop membranes or other protective structures for the embryo, except for the presence of a surrounding jelly. The egg is thus permeable allowing the exchange of gases and waste products by osmosis through the jelly capsule. The oxygen in the water diffuses through the jelly layer, across the membrane, through the perivitelline fluid and into the embryo, carbon dioxide and nitrogen waste (ammonia) move in the opposite direction also by diffusion. Consequently, these kind of eggs develop in water bodies[**1**, **5**].

In contrast to amphibian eggs, amniote eggs possess several innovations, including three additional embryonic layers: the amnion, the chorion and the allantois embryonic membranes. The amnion is a membrane forming a fluid-filled cavity that encloses the embryo. This transparent fluid where the embryo is suspended acts like a shock absorber and also provides protection against water loss and tissue adhesions[**6**]. The chorion is the outermost membrane around the embryo in reptiles, birds and mammals[**7**]. The allantois is an extra-embryonic membrane which together with the chorion are temporary respiratory organs as well as specialized structures for storing nitrogenous waste and converting ammonia into less toxic urea. Protecting the embryo from the toxic effects of its nitrogenous waste is regarded as a major innovation in the origin of amniotes' terrestrial eggs[**8**]. All these structures and also the yolk sac of the amniote egg are presented schematically in Figure 1.2.

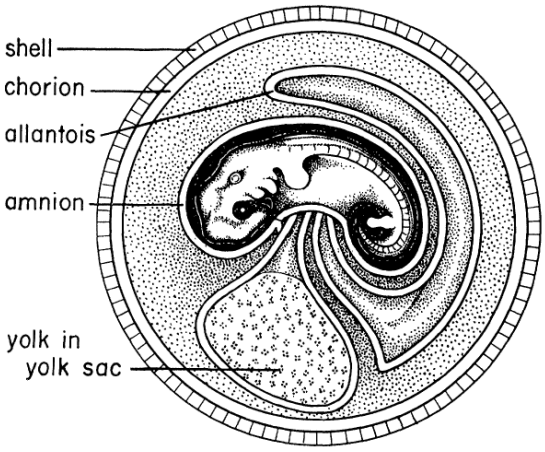

Figure 1.2: Schematic representation of the structure of an amniote egg, showing the growing embryo protected by the shell, the chorion, the amnion, the allantois and the yolk sac[**1**].

In most of the cases, amniote eggs are also bigger in size than amphibian eggs and the size of supporting embryonic layers increases disproportionately with respect to the size of the embryo. This implies that the fluids inside the egg would increase too. Moreover, physical support of the egg is even more important for eggs deposited in terrestrial environments where surface tension, in addition to gravity, would tend to deform the egg more than in the case of amphibian eggs laid on water. Consequently, the walls that contain the embryo should become thicker if the internal volume and the tension increase due to Laplace principle.

This previously discussed fact implies the replacement of the amphibian egg capsule by a fibrous shell membrane not to limite gas exchange between the embryo and its environment[**9**].

To sum up, the evolution from an amphibian egg to a more complex amniote egg involves a number of important innovations, including the modification of extraembryonic egg envelopes, the increase in the egg size and a stronger envelope (shell) covering the amniote egg.

A brief summary of the main differences between the amphibian and the amniote egg is shown in Table 1.1.

Structure	Amphibian egg	Amniote egg
Shell	No	Yes
Jelly-like substance	Yes	No
Yolk	Small	Big
Membranes (amnion, chorion and allantois)	No	Yes
Size	Small	Big

Table 1.1: Main differences between the structure of amniote egg and amphibian egg.

1.2 COMPARATIVE GENOMICS

Comparative genomics aims to understand the genomic basis of evolutionary change by looking at shared and specific genomic features across genomes from different species. Such differences can be in gene content (i.e., gene gains and losses) or their organization (e.g., synteny, chromosome evolution), among others. While performing such comparisons, it is of outtermost importance that evolutionary relationships among species are taken into account, i.e. the phylogenetic history or phylogeny. A phylogeny-aware comparative approach is a very powerful method to understand the genomic basis of innovations, because it allows to differentiate true evolutionary convergence from shared ancestry[**10**].

Darwin's theory of evolution states that all species have evolved from a common ancestor. The field of phylogenetics studies the evolutionary relationships among biological entities (different species, individuals of the same species and genes within a genome). A phylogenetic tree represents a hypothesis about how these entities evolved from a common ancestor. In a phylogenetic tree[**11**, **12**]:

1. Tips or terminal branches represent the species, individuals, or genes.
2. Internal branches represent ancestral lineages.
3. Nodes represent common ancestors of the tips (or branches) they give rise to.
4. The root provides the polarity of the tree, i.e. the directionality (in time).
5. Phylogenetic trees are usually bifurcating: a common ancestor gives rise to two separate entities (e.g., species, populations, paralogs).
6. The branching patterns (topology) reflects the evolutionary relationships of species, populations, or genes (branches) by common ancestry (nodes).
7. If present, branch lengths reflect the amount of evolutionary change, either in number of expected changes or in time units.
8. Two species, individuals or genes are more related to each other if they share a more recent common ancestor.

Comparative genomics has a central role in modern evolutionary biology. Moreover, comparative genomics is a powerful tool with several applications also in other fields such as medicine, forensics, epidemiology, drug design and agriculture[**11**, **12**].

Homology is a core concept in comparative genomics. For example, homology is used as a proxy for similar functions. Characterizing the function of a protein in vivo is complex and expensive and it does not scale up the currently available genomic data. Therefore, functional annotation is often extrapolated from homologous sequences in model organisms where their function has been experimentally established. Nevertheless, the inference of functional similarity from homology is not straightforward, in part because homologs in different species might not need to retain the same function[**13**]. Moreover, every gene can have multiple functions.

In this context, identifying homology relationships is at the core of comparative genomics. Specifically, differentiating among several types of homology (e.g. orthology and paralogy) is important. Gene duplication is considered one of the major sources of innovation in genomes. The concept of orthology was originally introduced to distinguish two kinds of evolutionary histories[**13**]:

1) Orthologs: homologous sequences originated through speciation events
2) Paralogs: homologous sequences originated by gene-duplication events.

It is generally assumed that upon duplication of a gene, one of the copies will retain the ancestral function and the other one can vary, either changing the pattern of expression across time or tissues (subfunctionalization) or acquire a new function (neofunctionalization)[**14**]. In the first case, paralogs will have different expression patterns, whereas in the second case they might have different functions. Therefore, orthologs are generally assumed to most likely retain the ancestral function.

1.3 BIOINFORMATIC METHODS

1.3.1 *The Ensembl database*

Ensembl[**15**] is one of several genome browsers for the retrieval of genomic information, specifically vertebrate genomes. This database supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl was launched in 1999 in response to completion of the Human Genome Project as a joint scientific project between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute. It also provides software to annotate genes, computes multiple alignments, predicts regulatory functions and collects disease data. Some of its tools include BLAST, BLAT, BioMart and the Variant Effect Predictor.

1.3.2 *Genome quality assessment with BUSCO*

BUSCO (Benchmarking Universal Single-Copy Orthologs[**16**]) is an open-source software quality assessment tool which provides quantitative measurements of the completeness of genomic data in terms of expected gene content. It identifies complete, single-copy, duplicated, fragmented and missing genes and enables like-for-like quality comparisons of different data sets employing ortholog sets from OrthoDB[**16**]. Sets of single-copy orthologs across multiple species are at the core of BUSCO, also known as "BUSCOs". Different sets of BUSCOs have been inferred for diverse groups of organisms, such as animals, arthropods or vertebrates. BUSCOs can be seen as an evolutionarily-informed expectation that these genes should be found as single-copy orthologs in any newly-sequenced genome. Because BUSCOs represent evolutionary conserved genes and are single-copy in most studied genomes for a particular lineage, the evolutionary expectation means that if a particular BUSCO cannot be identified in a new genome assembly, its absence is most probably due to errors in genome sequencing, assembly, or annotation[**16**, **17**].

Besides measuring the quality of genome assemblies for comparative genomics analyses, BUSCO has many other applications like building training sets gene predictors, controlling data quality and identifying reliable markers for large-scale phylogenomic and metagenomic studies[**17**]. Due to all of these applications, BUSCO has become established as a crucial bioinformatics tool.

1.3.3 *Sequence similarity searches*

BLAST (Basic Local Alignment Search Tool[**18**]) finds regions of local similarity between sequences. It allows to identify homologous sequences by detecting excess similarity, which is measured by the statistic known as e-value. Low e-values imply that two sequences share more similarity than would be expected by chance. In that case, the simplest explanation is that these sequences did not arise independently, but that they share a common ancestor. BLAST contains a number of algorithms to compare sequences of nucleotides or amino acids (known as queries) against databases

of nucleotide or amino acid sequences (known as databases). Specifically, BLASTP compares protein sequences to sequence databases that contain other proteins.

BLAST uses reliable statistical models to estimate whether an alignment similarity score would be expected by chance. Currently, protein databases contain tens of millions of sequences where the majority of them are unrelated to an individual query[**18**]. Thus, determining the distribution of scores expected by chance is described by the extreme value distribution1.1:

$$p(s \geq x) \leq 1 - e^{-e^{-x}} \tag{1.1}$$

where the score s has been normalized to correct for the scaling of the scoring matrix and the length of the sequences being compared.

To avoid these normalization issues, most similarity searching programs also provide a score in bits, which can be converted into a probability using the formula1.2:

$$p(b \geq x) \leq 1 - e^{-mn2^{-2}} \tag{1.2}$$

where m and n are the lengths of the two sequences being aligned and $p(b)$ is the probability of the score in a single pairwise alignment.

This search program reports the best scores after doing hundreds of thousands to tens of millions of comparisons. For this reason, BLAST reports the expected number of times the score would occur by chance, called expectation value or e-value, which depends also on database size.

Despite its high accuracy, BLAST can be computationally very demanding when using large sets of queries and databases, as often is the case in comparative genomic studies. To overcome this burden, faster software applications have been developed recently. One of such software is DIAMOND[**19**], which performs sequence similarity searches similarly to BLAST but at a fraction of the time. Benchmarking analyses have shown that DIAMOND is slightly less sensitive than BLAST, but still accurate[**19**].

1.3.4 *Protein clustering with MCL*

The MCL algorithm (Markov Cluster Algorithm[**20**]) is an unsupervised cluster algorithm for graphs which is described as fast and scalable. It was created by Stijn van Dongen and specifically designed for eukaryotic genomes. In bioinformatics, the MCL algorithm has been used to cluster homologous genes[**20**].

MCL is naturally described in matrix algebra. The MCL process generates a sequence of stochastic matrices (named Markov matrices) given some initial stochastic matrix and simulates flow alternating two simple algebraic operations on matrices. In the first operation, even index elements are obtained by expanding the previous element that coincides with normal matrix

multiplication. In the second operation, odd index elements are obtained by inflating the previous element given some inflation constant, which mathematically means a Hadamard power followed by a diagonal scaling. Inflation models the contraction of flow, it becomes thicker in regions of higher current and thinner in regions of lower current. These two operations can be summarized as matrix squaring (expansion) and rescaling the entries of a stochastic matrix to remain stochastic (inflation). The sequence of MCL elements from the process does not end until the elements converge to some specific kind of matrix, called the limit of the process. The heuristic underlying MCL predicts that the interaction of expansion with inflation will lead to a limit exhibiting cluster structure in the graph associated with the initial matrix. The number of clusters cannot and need not be specified in advance. A single parameter called inflation $-I$ controls the granularity of the output clustering. The granularity of the clusters defines how fragmented or aggregated the genes will be in the results. Usually, the inflation parameter is decided experimentally and depends on each dataset[**21**]. This algorithmic process is shown graphically in Figure 1.3.

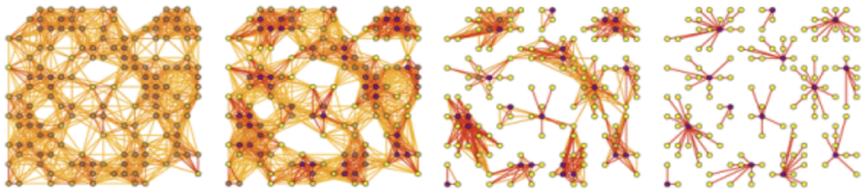

Figure 1.3: Scheme of the process how MCL operates[**20**].

Mathematically, the MCL process is described as follows. A MCL process is characterized by an infinite row of pairs (e_i, r_i), where e_i are integers greater than one, and r_i are real numbers greater than zero. An input matrix M yields an infinite number of matrices M_i by setting $M_1 = M$, defining the even-labeled iterands by setting M_{2i} to M_{2i-1} raised to the power e_i, and the odd-labeled iterands by $M_{2i+1} = \Gamma_{r_i}(M_{2i})$. The operator Γ_{r_i} transforms a column-stochastic matrix into another column-stochastic matrix by raising each entry to the power r_i and rescaling the result to be stochastic again[**21**].

To sum up, MCL transforms an input graph into an initial matrix suitable for starting the process, sets inflation parameters and does the MCL process. The result is then interpreted as clustering. MCL has been applied in a number of different domains, mostly in bioinformatics. One of the most important bioinformatic applications is the inference of sets of homologous genes into clusters.

1.3.5 *Automatic annotation with GO terms*

Gene Ontology was set up in 1998 by a consortium of researches studying the genomes of three species, fruitfly (*Drosophila melanogaster*), mouse (*Mus musculus*) and budding yeast (*Saccharomyces cerevisiae*)[22]. An ontology consists of a formal representation of concepts and the relationships between them within a given area which is structured as a directed acyclic graph. The Gene Ontology project provides an ontology of defined terms representing gene product properties, known as Gene Ontology (GO) terms. Each GO vocabulary has a term name, a unique alphanumeric identifier, a definition with cited sources and a namespace indicating the domain to which it belongs. These terms are designed to be species-neutral, and include terms applicable to prokaryotes and eukaryotes, single and multicellular organisms[22].

The Gene Ontology describes biological knowledge with respect to three aspects[22]:

- Cellular component: this feature refers to cellular anatomy and describes parts of a cell where a gene product performs its function, either cellular compartments or stable macromolecular complexes of which they are parts.
- Biological process: this characteristic involves multiple molecules accomplishing larger processes, for instance DNA repair or signal transduction, which are essential for cells, tissues, organs, and organisms.
- Molecular function: this attribute is related to molecular-level activities which are carried out by gene products. They generally correspond to activities that can be performed by individual gene products (a protein or RNA), but some of them are completed by molecular complexes composed of multiple gene products.

One of the applications of GO terms is to functionally characterize sets of genes in non-model organisms. This is based on the principle that homologous sequences from different species will share the same or similar functions, and thus GO annotations from model organisms can be transferred to other species based on sequence homology. Among the many possible ways of studying GO annotations, one of the most common ones is to perform enrichment tests. This analyses test for the overrepresentation of GO terms in a set of annotated genes. Several software applications have been developed to perform enrichment tests of GO terms, for instance topGO[23]. The most common statistical test is Fisher's exact test.

2 METHODS

This section describes the steps for the selection of high-quality genomes for comparative genomics with BUSCO, the estimation of ancestral and novel sets of proteins in the ancestor of amniotes and other major groups of tetrapods using the Phylogenetic Aware Parsing Script pipeline (PAPS; Paps and Holland 2018[**24**]), and the functional annotation of genes that originated in the ancestor of amniotes.

2.1 SELECTION OF SPECIES

In this study, the Ensembl releases 95 and 96 were used for selecting the species to be analyzed. Ensembl contains high-quality genomes from vertebrates, including birds, reptiles, mammals and fishes. Despite the aim of being representative of the existing diversity, the representation of different vertebrate lineages in Ensembl is necessarily biased, as it reflects the current bias in sequenced genomes. Of all the available genomes, we chose 108 species from release 95 and 33 more from release 96, after excluding duplicated genomes for the same genus. New genomes that appeared in release 96 were later incorporated because they included several relevant species, including several previously unrepresented reptiles. For the total of 141 genomes, the annotated set of peptides (*.pep.all.fa) were downloaded from the ftp site of Ensembl:

- ftp://ftp.ensembl.org/pub/release-95/fasta/
- ftp://ftp.ensembl.org/pub/release-96/fasta/

The selection list of the original species alphabetically ordered for each release is shown below in Tables 2.1 and 2.2. In Figures 2.1, 2.2, 2.3 and 2.4, the species are classified into major evolutionary groups to show how these are represented by the currently available species.

Anser brachyrhynchus	*Junco hyemalis*	*Parus major*
Apteryx owenii	*Lepidothrix coronata*	*Piliocolobus tephrosceles*
Bison bison bison	*Lonchura striata domestica*	*Pogona vitticeps*
Calidris pugnax	*Manacus vitellinus*	*Prolemur simus*
Castor canadensis	*Marmota marmota marmota*	*Salvator merianae*
Chelonoidis abingdonii	*Melopsittacus undulatus*	*Serinus canaria*
Coturnix japonica	*Meriones unguiculatus*	*Spermophilus dauricus*
Cricetulus griseus picr	*Neovison vison*	*Theropithecus gelada*
Crocodylus porosus	*Notechis scutatus*	*Urocitellus parryii*
Cyanistes caeruleus	*Nothoprocta perdicaria*	*Ursus maritimus*
Dromaius novaehollandiae	*Numida meleagris*	*Zonotrichia albicollis*

Table 2.1: Original selection of species from Ensembl 96.

Acanthochromis polyacanthus	*Gambusia affinis*	*Oryzias latipes*
Ailuropoda melanoleuca	*Gasterosteus aculeatus*	*Otolemur garnettii*
Amphilophus citrinellus	*Gopherus agassizii*	*Ovis aries*
Amphiprion percula	*Gorilla gorilla*	*Pan troglodytes*
Anabas testudineus	*Heterocephalus glaber female*	*Panthera pardus*
Anas platyrhynchos	*Hippocampus comes*	*Papio anubis*
Anolis carolinensis	*Homo sapiens*	*Paramormyrops kingsley*
Aotus nancymaae	*Ictalurus punctatus*	*Pelodiscus sinensis*
Astyanax mexicanus	*Ictidomys tridecemlineatus*	*Periophthalmus magnuspinnatus*
Bos taurus	*Jaculus jaculus*	*Peromyscus maniculatus bairdii*
Callithrix jacchus	*Kryptolebias marmoratus*	*Phascolarctos cinereus*
Canis familiaris	*Labrus bergylta*	*Poecilia formosa*
Capra hircus	*Latimeria chalumnae*	*Pongo abelii*
Carlito syrichta	*Lepisosteus oculatus*	*Procavia capensis*
Cavia porcellus	*Loxodonta africana*	*Propithecus coquereli*
Cercocebus atys	*Macaca nemestrina*	*Pteropus vampyrus*
Chinchilla lanigera	*Mandrillus leucophaeus*	*Pygocentrus nattereris*
Chlorocebus sabaeus	*Mastacembelus armatus*	*Rattus norvegicus*
Choloepus hoffmanni	*Meleagris gallopavo*	*Rhinopithecus bieti*
Chrysemys picta bellii	*Mesocricetus auratus*	*Sarcophilus harrisii*
Colobus angolensis palliatus	*Microcebus murinus*	*Scleropages formosus*
Cynoglossus semilaevis	*Microtus ochrogaster*	*Scophthalmus maximus*
Cyprinodon variegatus	*Mola mola*	*Seriola dumerili*
Danio rerio	*Monodelphis domestica*	*Sorex araneus*
Dasypus novemcinctus	*Monopterus albus*	*Sphenodon punctatus*
Dipodomys ordii	*Mus musculus*	*Stegastes partitus*
Echinops telfairi	*Mustela putorius furo*	*Sus scrofa*
Equus caballus	*Myotis lucifugus*	*Taeniopygia guttata*
Erinaceus europaeus	*Nannospalax galili*	*Takifugu rubripes*
Esox lucius	*Nomascus leucogenys*	*Tetraodon nigroviridis*
Felis catus	*Notamacropus eugenii*	*Tupaia belangeri*
Ficedula albicollis	*Ochotona princeps*	*Tursiops truncatus*
Fukomys damarensis	*Octodon degus*	*Vicugna pacos*
Fundulus heteroclitus	*Oreochromis niloticus*	*Vulpes vulpes*
Gadus morhua	*Ornithorhynchus anatinus*	*Xenopus tropicalis*
Gallus gallus	*Oryctolagus cuniculus*	*Xiphophorus maculatus*

Table 2.2: Original selection of species from Ensembl 95.

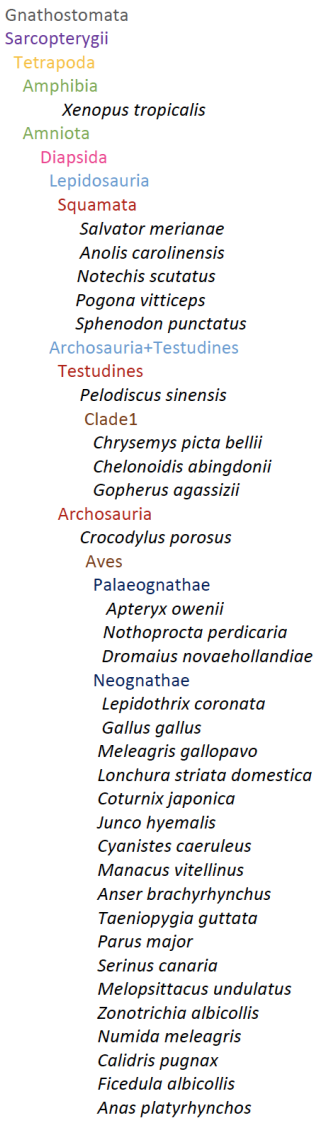

Figure 2.1: Evolutionary classification levels considered represented in different colors. Note that clade1 was only defined to be able to infer turtle's (turtles1) ancestral and novel gene sets.

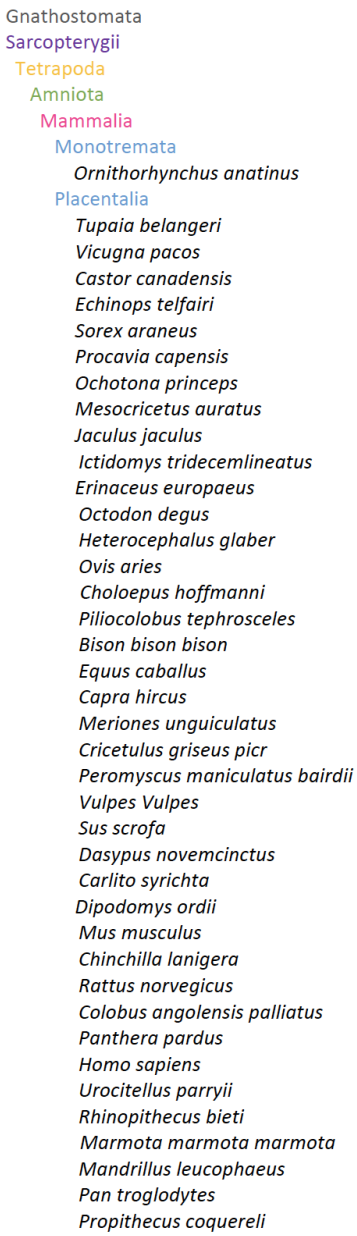

Figure 2.2: Evolutionary classification levels considered represented in different colors. Note that clade1 was only defined to be able to infer turtle's (turtles1) ancestral and novel gene sets.

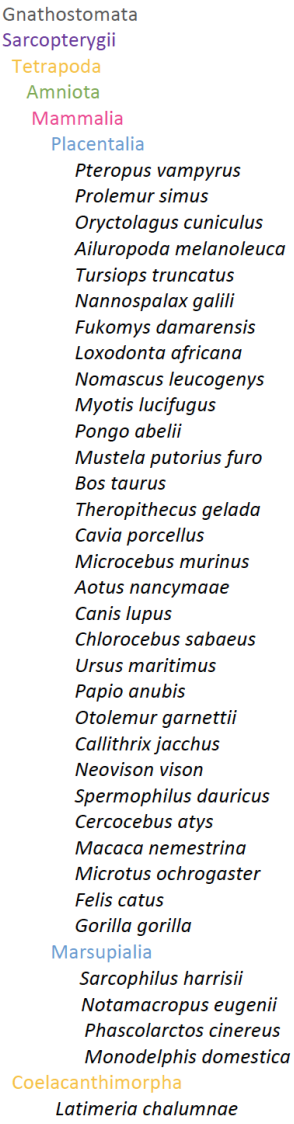

Figure 2.3: Evolutionary classification levels considered represented in different colors. Note that clade1 was only defined to be able to infer turtle's (turtles1) ancestral and novel gene sets.

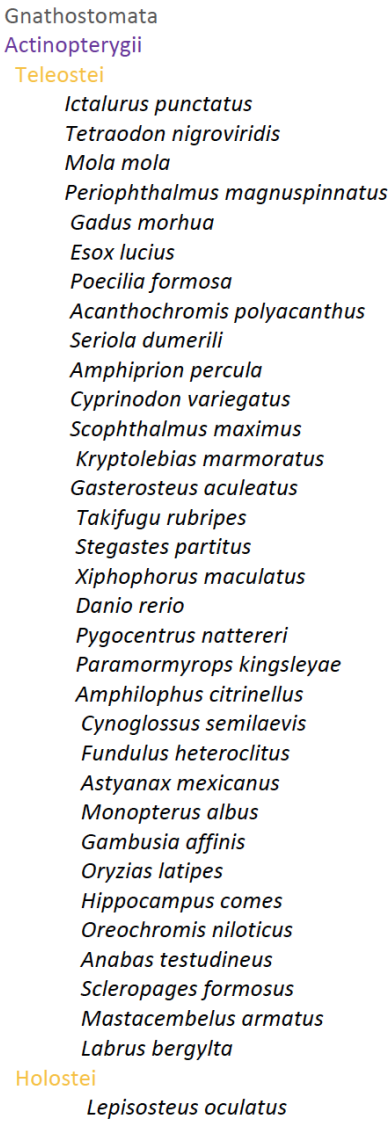

Figure 2.4: Evolutionary classification levels considered represented in different colors. Note that clade1 was only defined to be able to infer turtle's (turtles1) ancestral and novel gene sets.

2.2 GENOME QUALITY ASSESSMENT WITH BUSCO

The assessment of the quality of genome assemblies is an essential step in order to be able to discard those that are of low quality. Low quality genomes originate many problems when making comparative genomics and thus it is better to exclude them as soon as possible (Milinkovitch et al. 2010[**25**]). The removal of low-quality genomes will also help to reduce the computational burden of downstream analyses. BUSCO provides quantitative measures of the completeness of genome assemblies in terms of expected content of single-copy orthologs derived from OrthoDB v.9[**16**]. The set of orthologs used by BUSCO needs to be tailored to evolutionary groups being studied. In our case, the vertebrate dataset (`vertebrataodb9`) was used, which contains a total of 3023 BUSCOs.

For each of the original selected species, the BUSCO software was run as follows:

```
python2.7 /gpfs/resapps/BUSCO/3.0.2/scripts/runBUSCO.py --in
speciesfilename.pep.all.fa -l vertebrataodb9 -m proteins --out
        outputnamefile --cpu 10 --evalue 1e-5
```

where $--in$ provided the genome to be evaluated; $-l$ vertebrataodb9 was the reference set of BUSCOs; $-m$ proteins was the type of analysis to run for annotated gene sets or proteins; $--cpu$ 10 was the number of threads/cores used; and $--$evalue 10^{-5} was the e-value cutoff for BLAST searches.

A genome assembly was considered to be of high quality whenever it had over 90 % of complete genes of the 3023 single-copy orthologs (BUSCOs) in the test set. An even more stringent threshold of 95 % was not considered because several phylogenetically important species (of relevance for the downstream comparative genomics analyses) would have been discarded (see Results and Discussion 3.1). Therefore, not only the proportion of complete BUSCOs was used as a criterion, but also the phylogenetic position of the species.

2.3 PHYLOGENETIC AWARE PARSING SCRIPT

Phylogenetic Aware Parsing Script or PAPS[**24**, **26**] (available on https://github.com/ PapsLab/PhylogeneticAwareParsingScript) is a pipeline that produces lists of homologous groups (HG) using sequence similarity (e.g. BLASTP) and clustering (e.g. MCL), taking the evolutionary relationships of the species into account. The main goal of the PAPS pipeline is to infer the patterns of gene gains and losses along a phylogeny. The pipeline is composed of three perl scripts. In the last step, the user can introduce search criteria to obtain sets of HGs associated with a given node and custom patterns of presence/absence across evolutionary groups.

2.3.1 *Preparation of the proteome files*

A first step to run the PAPS pipeline is to include a short label representing the species name. This label should be unique and be included at the beginning of the sequence name (in the fasta header; just after the '>' symbol). These labels were chosen so that they are representative of the species' names (all the labels and an explanation of the fasta format are in Appendices A.1 and A.2, respectively). The original description of sequences was also simplified, keeping only Ensembl's unique protein IDs. The process how all these files were modified is shown in Figure 2.5 for one of them.

```
[lavinm@altamira1 ~]$ head -n3 Acanthochromis_polyacanthus.ASM210954v1.pep.all.fa
>ENSAPOP00000008786.1 pep primary_assembly:ASM210954v1:MVNR01004827.1:2681:2737:1 gene:ENSAPOG00000010967.1 transcript:ENSAPOT00000002662.1 gene_biotype:IG_J_gene transcript_biotype:IG_J_gene
SYFDYWGKGTQVTVTSGK
>ENSAPOP00000010247.1 pep primary_assembly:ASM210954v1:MVNR01007420.1:817:873:1 gene:ENSAPOG00000012658.1 transcript:ENSAPOT00000000162.1 gene_biotype:IG_J_gene transcript_biotype:IG_J_gene
[lavinm@altamira1 ~]$ sed -i -E '/>/ s/ .+//g' Acanthochromis_polyacanthus.ASM210954v1.pep.all.fa
[lavinm@altamira1 ~]$ head -n3 Acanthochromis_polyacanthus.ASM210954v1.pep.all.fa
>ENSAPOP00000008786.1
SYFDYWGKGTQVTVTSGK
>ENSAPOP00000010247.1
[lavinm@altamira1 ~]$ sed -i '/>/ s/>/>Apol_/g' Acanthochromis_polyacanthus.ASM210954v1.pep.all.fa
[lavinm@altamira1 ~]$ head -n3 Acanthochromis_polyacanthus.ASM210954v1.pep.all.fa
>Apol_ENSAPOP00000008786.1
SYFDYWGKGTQVTVTSGK
>Apol_ENSAPOP00000010247.1
```

Figure 2.5: Preparation of the header of one set of predicted proteins.

2.3.2 *Creation of DIAMOND databases*

Once the headers are modified in all the files, each containing the set of predicted proteins for a species, all the files were concantenated into a single file, which was called "`allproteomesdb`".

Using this file, a database was created containing all the final species set of predicted proteins ("`allproteomesdb`"). This step prepares the database for the subsequent sequence similarity searches by DIAMOND[**19**]. The following command was used:

```
diamond makeblastdb --in allproteomesdb --db allproteomesdb
```

where --in provided the input and --db provided the name of the output to be created by DIAMOND.

2.3.3 *Searching protein sequence similarities with DIAMOND*

Then, an all versus all sequence similarity search was done to identify homologous proteins among all the species. In practice, the database containing all sets of predicted proteins was searched using all individual genomes as queries using DIAMOND. For its higher computational efficiency, the software DIAMOND[**19**] was used instead of BLASTP. An e-value threshold of 10^{-5} was chosen following Paps and Holland[**24**]. The command used for each proteome was as follows, providing the query genome, the "`allproteomesdb`"database and additional options for output name and format:

```
diamond blastp --query speciesfilename.pep.all.fa --db
  allproteomesdb --evalue 0.00001 --outfmt 6 --out
              blastoutputspeciesfilename
```

After all the similarity searches were run, the outputs obtained for all the genomes were merged in a single file called "`allblastpoutput`".

2.3.4 *Gene clustering in homologous groups with MCL*

Using the sequence similarities inferred by DIAMOND, MCL was used to cluster genes from different species into HGs. To prepare the DIAMOND output for MCL[**20**, **21**], a dependency called mcxdeblast was used with the following command:

```
          mcxdeblast --m9 --line-mode=abc
    --out=mcxdeblastallblastpoutput allblastpoutput
```

Then, the MCL clustering was performed as follows, using an inflation value *-I* of 2.0, following [**24**] and [**27**]:

```
mcl mcxdeblastallblastpoutput -I 2 --abc -o mclallblastpoutput
```

After, "`MCLrowcounter.pl`"script which is in PAPS pipline parsed the output of MCL called "`mclallblastpoutput`"to produce a taxonomic occupancy table by placing in the same directory this script, "`mclallblastpoutput`"plus "`allproteomesdb`". MCL row counter perl script runs with these two previous files as input. The result file has a HG in each row and one species per column. Each number of each cell indicates how many sequences of that especies are present in that HG. MCL row counter script must have been modified by introducing the labels that were written in each header of each species genomes file. These labels had been introduced in the array in line 81 of this perl script.

2.3.5 *Preparation of the PAPS script*

The output of MCL ("`mclallblastpoutput`") was parsed with the script "`MCLrowcounter.pl`", which is included into the PAPS pipeline to produce a taxonomic occupancy table. In order to do this, the script, MCL results ("`mclallblastpoutput`") and the original set of predicted proteins ("`allproteomesdb`") were placed into the same directory. This script uses the information of the sequence labels for identifying the species each sequence belongs too (this was appended earlier to the fasta headers; see A.1). Prior to execution, this script was modified (line 81) to hard-code the species specific labels. The resulting output file has one HG per row and one species per column and numbers at cells indicate how many sequences of a given species are present for a given HG.

A second perl script named "`CreateDBs.pl`", within the PAPS pipeline, was used to speed up the subsequent steps. Following the instructions by the authors, lines 9, 10 and 11 were modified to match our file names. To allow the next step, the permissions of the resulting database files were changed to make them available to all users.

The last step in the PAPS pipeline is the perl script named "`PAPS.pl`". In order to make it work with our data, the labels identifying each species were hard-coded in lines 35, 229, 471 and 499. Lines 12 and 13 were also modified to match our input and output file names. An additional important modification of the script was the customization of the multidimensional data structure containing the information of the evolutionary relationships among species (i.e. the phylogeny). In the script, this is done using a hash of hashes named "$spp"that needs to be modified to accomodate the species and the phylogeny being used. In our case, we used the species phylogeny provided by Ensembl. This tree structure is specified in line 590 and following. It is important that all species have the same number of classification levels in the hash. Empty classification levels ({' '}) can be used but they cannot be empty in all. Also, in the hash of hashes, each species needs to be assigned a value corresponding to an index of its position in the hash, starting from 0. The classification levels that were considered are shown in Figure 2.6.

- Gnathostomata
- Actinopterygii or Sarcopterygii
- Holostei, Teleostei, Coelacanthimorpha or Tetrapoda
- Amniota or Amphibia
- Diapsida or Mammalia
- Archosauria and Testudines, Lepidosauria, Marsupialia, Monotremata or Placentalia
- Archosauria, Testudines or Squamata
- Aves or Turtles (Clade1)
- Neognathae or Palaeognathae
- Species

A diagram with the different classifcation levels is shown in Figure 2.6.

Figure 2.6: Evolutionary levels considered represented in different colors.

2.3.6 *Inferring ancestral and novel genes with PAPS*

The last step in the PAPS pipeline is the perl script "`PAPS.pl`", which provided the correct input files (see above), allows for interactive searches for specific patterns of HG distribution across the phylogeny. When the PAPS script is executed, a command prompt will ask the user for search criteria about the presence or absence of the HG in different clades of interest.

In this study, we followed Paps and Holland[**24**, **26**] in the definition of four types of HGs:

- Ancestral HGs: the HGs present in the last common ancestor of a given clade. These might be also present in other clades.
- Ancestral Core HGs: a subset of Ancestral HGs, with the constraint that they must be present in all species, or all but one. These aim to represent essential HGs for a particular clade.
- Novel HGs: the HGs present in the last common ancestor of a clade but not in the outgroups (i.e. rest of clades). These are a subset of Ancestral HGs. Novel HGs are defined as present in at least one species from the in-group lineage positioned as sister group to the rest of the clade and in at least one species from of the rest of the clade; e.g. a novel HG in Sarcopterygii must be present in one species each of Coelacanthimorpha and Tetrapoda (see Fig. 3.4).
- Novel Core HGs: a subset of novel HGs present in every representative species within the clade or all but one. These are a subset of Novel HGs and aim to represent essential Novel HGs for a particular clade.

These categories of HGs were searched for a number of representative vertebrate clades using the syntax of the PAPS pipeline, as shown below:

- Sarcopterygii
  - Ancestral: `Tetrapoda-atleast1 Coelacanthimorpha-atleast1`
  - Ancestral Core: `Sarcopterygii-minus1`
  - Novel: `Tetrapoda-atleast1 Coelacanthimorpha-atleast1 outgroup-absent`
  - Novel core: `Sarcopterygii-minus1 outgroup-absent`
- Tetrapoda
  - Ancestral: `Amphibia-atleast1 Amniota-atleast1`
  - Ancestral Core: `Tetrapoda-minus1`
  - Novel: `Amphibia-atleast1 Amniota-atleast1 outgroup-absent`
  - Novel core: `Tetrapoda-minus1 outgroup-absent`
- Amniota
  - Ancestral: `Diapsida-atleast1 Mammalia-atleast1`
  - Ancestral Core: `Amniota-minus1`
  - Novel: `Diapsida-atleast1 Mammalia-atleast1 outgroup-absent`
  - Novel core: `Amniota-minus1 outgroup-absent`

- Diapsida
  - Ancestral: `Lepidosauria-atleast1 Archosauria+Testudines-atleast1`
  - Ancestral Core: `Diapsida-minus1`
  - Novel: `Lepidosauria-atleast1 Archosauria+Testudines-atleast1 outgroup-absent`
  - Novel core: `Diapsida-minus1 outgroup-absent`
- Archosauria+Testudines
  - Ancestral: `Archosauria-atleast1 Testudines-atleast1`
  - Ancestral Core: `Archosauria+Testudines-minus1`
  - Novel: `Archosauria-atleast1 Testudines-atleast1 outgroup-absent`
  - Novel core: `Archosauria+Testudines-minus1 outgroup-absent`
- Archosauria
  - Ancestral: `Crocodylusporosus-atleast1 Aves-atleast1`
  - Ancestral Core: `Archosauria-minus1`
  - Novel: `Crocodylusporosus-atleast1 Aves-atleast1 outgroup-absent`
  - Novel core: `Archosauria-minus1 outgroup-absent`
- Testudines
  - Ancestral: `Pelodiscussinensis-atleast1 turtles1-atleast1`
  - Ancestral Core: `Testudines-minus1`
  - Novel: `Pelodiscussinensis-atleast1 turtles1-atleast1 outgroup-absent`
  - Novel core: `Testudines-minus1 outgroup-absent`
- Aves
  - Ancestral: `Neognathae-atleast1 Palaeognathae-atleast1`
  - Ancestral Core: `Aves-minus1`
  - Novel: `Neognathae-atleast1 Palaeognathae-atleast1 outgroup-absent`
  - Novel core: `Aves-minus1 outgroup-absent`
- Lepidosauria
  - Ancestral: `Sphenodonpunctatus-atleast1 Squamata-atleast1`
  - Ancestral Core: `Lepidosauria-minus1`
  - Novel: `Sphenodonpunctatus-atleast1 Squamata-atleast1 outgroup-absent`
  - Novel core: `Lepidosauria-minus1 outgroup-absent`
- Mammalia
  - Ancestral: `Monotremata-atleast1 Placentalia-atleast1`
  - Ancestral Core: `Mammalia-minus1`
  - Novel: `Monotremata-atleast1 Placentalia-atleast1 outgroup-absent`
  - Novel core: `Mammalia-minus1 outgroup-absent`

Each of the search queries produced four output files[**26**]:

- ... MCLannotatedgenes.out: it contains a list of HGs, and within each HGs a list of included genes; one species per line.
- ... MCLcolumnsparsed.out: occupancy table for each HG (rows) and species (columns), indicating the number of genes for a particular species in a given HG.
- ... MCLgenesIDs.out: it contains a tab-separated list of sequences included in each HG; one HG per row.
- ...HGstaxanames.out: it contains a list of taxa present in each HG.

Using the output of PAPS, the number of ancestral, ancestral core, novel, and novel core HGs for relevant vertebrate clades were obtained. The ancestral and novel genes for amniotes were further analyzed.

2.4 REMOVAL OF FALSE POSITIVES

The absence of invertebrates or unicellular organisms in the source dataset likely introduced false positives among the inferred sets of novel genes. Therefore, a first step prior to functional annotation was to identify and remove false positives from the set of amniote novel genes. To do so, the strategy was to use a similarity search and eliminate all HGs containing at least one sequence with significant similarity to any other sequence in NCBI's NR (non-redundant) protein database. This was done on the sets of 3865 and 8 amniote novel and novel core HGs, respectively. In practice, we first extracted all the sequences from the novel sets using their sequence identifiers.

Then, a modified version of NR was prepared by removing all sequences belonging to any genus used in our comparisons (to avoid self-hits). DIAMOND was used to perform the sequence similarity search using the following command:

```
diamond blastp --query Amniotanovelseq.fa --db nrwousedgenera
            --evalue 0.00001 --outfmt 6 --out
          Amniotanovelvsnrwousedgenera --threads 20
```

All hits with a e-value of 10^{-5} or less were considered significant (shown in the first row of DIAMOND's output). The sequence identifiers of significant hits were extracted and used to find out HGs that contained at least one of the significant hits. This step was done with a custom perl script ("`searchHGwithFasePos.pl`").

All 8 novel core HGs contained at least one false positive and were discarded for further analyses. From the total of 3865 novel HGs, 3781 contained at least one false positive and thus were excluded from further analyses, whereas 84 HGs contained no false positives. The set of 84 HGs was thus considered to genuinely represent the set of novel HGs in the ancestor of amniotes and were further studied in detail.

2.5 ANNOTATION WITH GENE ONTOLOGY TERMS

In order to functionally characterize the set of amniote novel HGs, these were annotated with GO terms and the overrepresented functions inferred with respect to the set of amniotes' ancestral set of HGs. For the purpose of the current study, the GO annotations referring to biological process were taken into account.

2.5.1 *Obtaining GO information from Ensembl*

GO information for all used 123 genomes were downloaded from Ensembl using the BioMart data mining tool (www.ensembl.org/info/data/biomart/biomartrestful.html#biomartperlapi). In practice, a sample xml query was created to contain the desired information, and further modified to access the information from all 123 genomes. All queries to obtain GO annotations were collected in the script "`queryGOtermsfromensembl.sh`". The downloaded information contained all GO annotations available for all genomes. From this information, the GO annotations of the genes inferred to be presented in the ancestor of amniotes were extracted (84 HGs and 14901 genes in total).

2.5.2 *Enrichment analysis with topGO and summary of results with REVIGO*

In order to infer the overrepresented functions among amniotes' novel HGs, these were compared with the set of amniote ancestral HGs as a baseline using Fisher's exact test. The software topGO was used, which was fed with the annotations of amniote ancestral HGs (Ensembl geneIDs and their associated GO terms) and a list of genes of interest to be used as query (novel HGs). The significance threshold was set at $p<0.01$. The results were summarized with the REVIGO webserver (http://revigo.irb.hr), which uses each GO term and its associated p-value from Fisher tests and generates three plots: a scatterplot, an interactive map and a treemap.

2.6 FLOWCHART OF THE METHODS

A flowchart to summarize all the steps followed is shown in Figure 2.7.

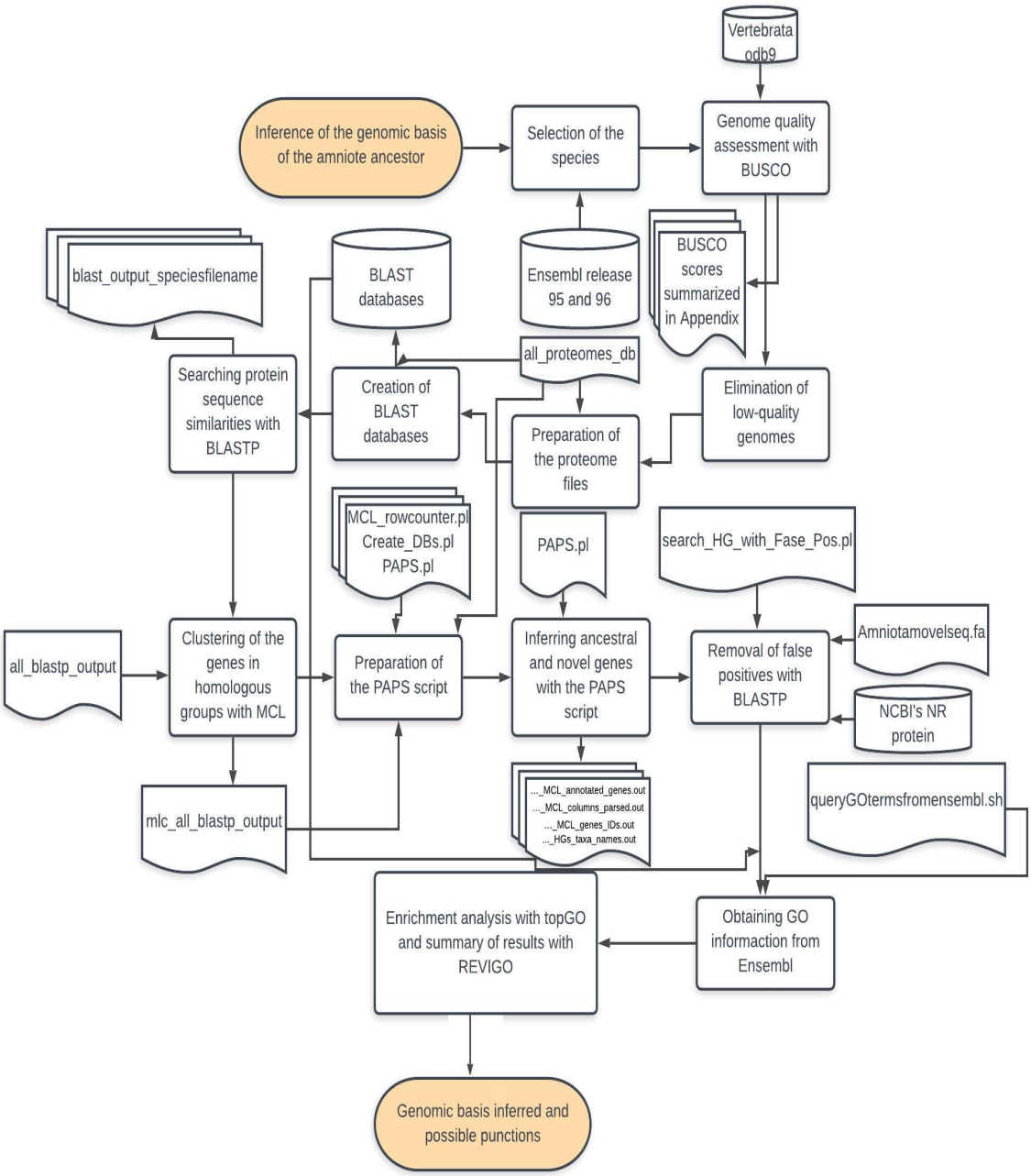

Figure 2.7: Flowchart of the steps followed, databases used and generated output.

3 RESULTS AND DISCUSSION

This section presents the obtained results and discusses them in the context of the proposed biological questions.

3.1 BUSCO ASSESSMENTS OF VERTEBRATE GENOMES

A graphical summary of the qualities of all the tested genome assemblies can be seen in Figures 3.2 and 3.3, including the proportion of complete, fragmented and missing BUSCOs. Overall, most assemblies obtained relatively high proportion of complete BUSCOs, which is expected given the aim of Ensembl to contain only high-quality genomes for comparative genomics[**15**]. For example, 122 and 85 out of the 141 genomes recovered $\geq$90 % and $\geq$95 % complete BUSCOs, respectively. According to our criterion of using completeness of single-copy orthologs as a proxy for high assembly quality (see Materials and Methods 2.2), 123 species (including *Ornithorhyncus anatinus*, see below) out of 141 with $\geq$90 % of complete BUSCOs were used for subsequent steps. This meant that 15 and 3 assemblies were dismissed from Ensembl releases 95 (Table 3.1) and 96 (Table 3.2). In Figure 3.1, the discarded species are shown together with their classification levels.

Choloepus hoffmanni	*Jaculus jaculus*	*Procavia capensis*
Dipodomys ordii	*Mesocricetus auratus*	*Sorex araneus*
Echinops telfairi	*Notamacropus eugenii*	*Tetraodon nigroviridis*
Erinaceus europaeus	*Ochotona princeps*	*Tupaia belangeri*
Gadus morhua	*Periophthalmus magnuspinnatus*	*Vicugna pacos*

Table 3.1: Eliminated species from Ensembl 95 alphabetically ordered.

Castor canadensis	*Notechis scutatus*	*Nothoprocta perdicaria*

Table 3.2: Eliminated species from Ensembl 96 alphabetically ordered.

Despite the platypus (*Ornithorhyncus anatinus*), having 76 % complete BUSCOs it was retained for subsequent steps given its key phylogenetic position as only representative of monotremes. Also, a more stringent threshold of 95 % complete BUSCOs was not used because this would have meant to exclude several species with key phylogenetic positions, such as representatives of sarcopterygian fish (*Latimeria chalumnae*), amphibians (*Xenopus tropicalis*), and reptiles (*Anolis carolinensis* and *Pelodiscus sinensis*) all of which were the only or one of the few representatives of their evolutionary lineages.

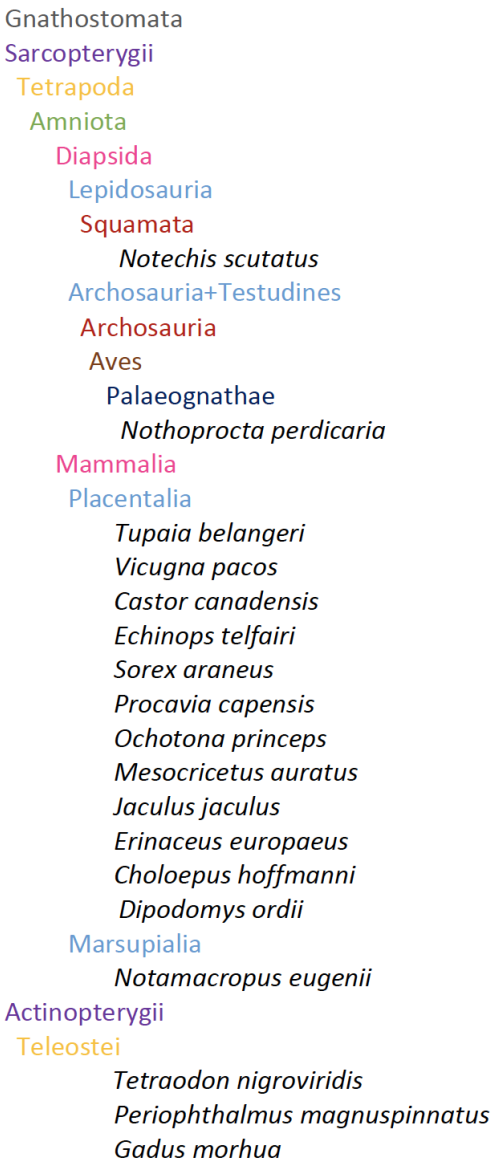

Figure 3.1: Low-quality genomes eliminated indicating their evolutionary affinities. Classification levels represented in different colors.

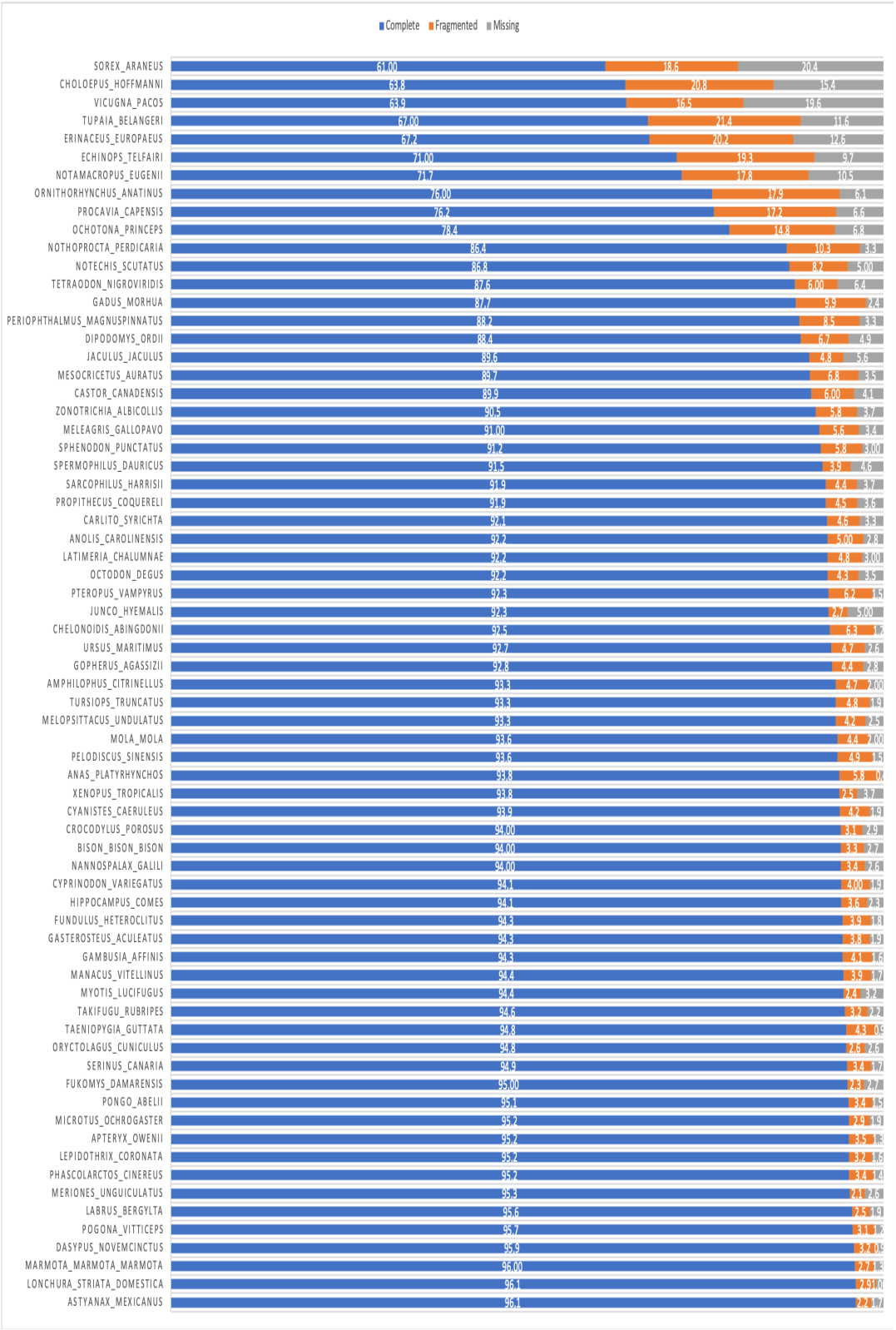

Figure 3.2: Proportion of complete (blue), fragmented (orange), and missing (grey) BUSCOs for the analyzed genome assemblies.

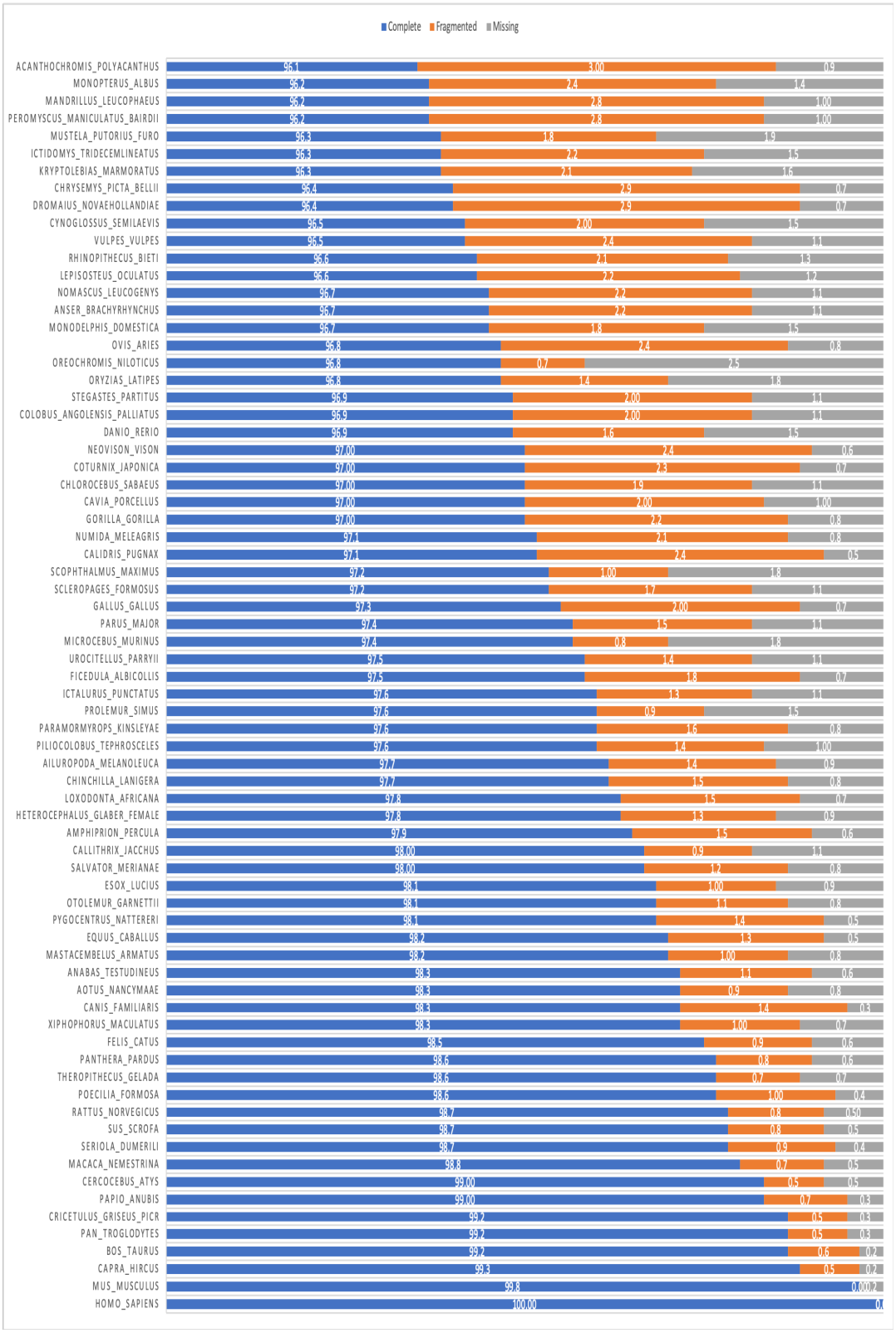

Figure 3.3: Proportion of complete (blue), fragmented (orange), and missing (grey) BUSCOs for the analyzed genome assemblies.

A second criterion to evaluate genome assemblies with BUSCO might be the proportion of single-copy and duplicated BUSCOs. This assumes that a higher proportion of single-copy BUSCOs and lower proportion of duplicated ones reflects a more contiguous assembly (i.e. of higher quality), whereas a genome with a high proportion of duplicated BUSCOs might be seen as more fragmented. However, a higher proportion duplicated BUSCOs did not correspond with a lower completeness scores and the genome with the highest number of duplicated BUSCOs was in fact the human genome, probably the best assembly available. The human genome contains 62.2 % duplicated BUSCOs, similarly to other hominids (e.g. *Gorilla gorilla* has 45.4 %) (Appendix A.3). This probably reflects the higher effort and resources used in improving the annotation of such genomes, and thus the proportion of duplicated BUSCOs was not used as criterion to assess genome contiguity.

3.2 HOMOLOGY GROUPS IN THE MAIN VERTEBRATE LINEAGES

Even though the main objective of this study was not to infer the evolution of genomic novelty among main vertebrate clades, the application of our pipeline allowed us to identify some interesting patterns. The inferred sets of ancestral genes for all groups range from 12153 to 18818, which accords well with the expected number of vertebrate genomes (see Fig. 3.4).

Gene innovation is inferred to be high during the early diversification of vertebrate lineages, in particular, in the origin of Sarcopterygii, Tetrapoda, Amniota, and Diapsida (see Fig. 3.4). These steps correspond to important changes in the morphology and lifestyle, including water-to-land transition in Tetrapods and full terrestrialization in Amniota [**3**, **27**].

More restricted clades (i.e. more recent in evolutionarily terms) had higher proportion of ancestral core genes, probably reflecting more homogeneous genomes, morphology, and lifestyles. For example, 183 ancestral core HGs were inferred for Sarcopterygii and 2165 for Aves (see Fig. 3.4).

There are also some limitations in this analysis. As suggested by our identification of false positives of novel HGs in amniotes, the inferred numbers for other clades might also contain a number of false positives, and thus the mentioned patterns should be taken with caution.

Amniotes were initially inferred to have 3865 novel HGs and 8 novel core HGs. These are the HGs that were analyzed in detail in the rest of our work.

3.3 ANNOTATION OF NOVEL GENES IN AMNIOTES

3.3.1 *False positives*

The identification of false positives indicated that 3781 out of 3865 amniote novel HGs contained false positives, i.e., at least one gene in these HGs showed significant similarity to sequences from other species not in the test set. This likely indicates that the raw numbers of HGs inferred with PAPS might be inflated. The reason for the high proportion of false positives is probably that non-vertebrates were not included in the test set, while these represent the vast majority of the diversity (invertebrates, unicellular eukar-

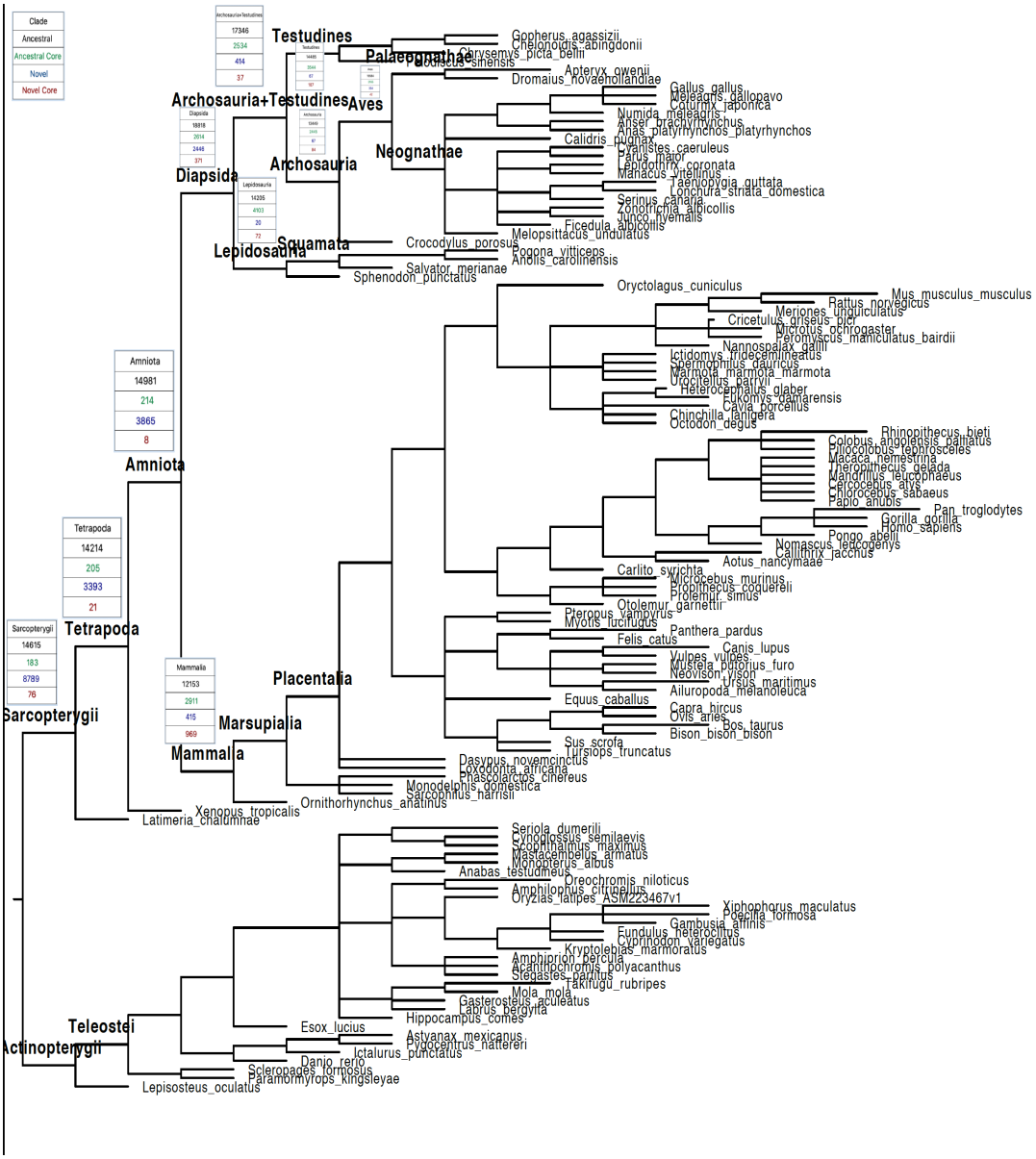

Figure 3.4: Sets of Ancestral (black), Ancestral Core (green), Novel (blue) and Novel Core HGs (red) inferred for representative vertebrate clades plotted onto a consensus phylogenetic tree (obtained from Ensembl). Main clade names are also highlighted.

yotes and prokaryotes). To reduce this bias in the set of novel amniote genes that are central to this study, we aimed to reduce the set of false positives by identifying and removing sequences that were homologous to other species outside the test set. A total of 84 Ancestral HGs were free of false positives.

3.3.2 *GO terms and results for enriched functions*

The genes included in those 84 novel HGs were characterized by obtaining their GO terms and performing an enrichment test against the set of amniotes' ancestral HGs. The result is a set of 213 GO terms that are enriched in the novel HGs (with $p<0.01$). This 213 GO terms were summarized with the REVIGO webserver by using each GO term and its associated p-value. The results are shown in Figures 3.5, 3.6 and 3.7.

The set of genes included in the 84 novel HGs are enriched in the following putative functions (inferred from GO terms as proxy):

Vitamin D biosynthesis, e.g. regulation of calcidiol 1-monooxygenase (enzyme involved in the modification of calciodiol into calciotriol, an active form of Vitamin D) and general Vitamin D biosynthesis regulation. Vitamin D is a fat-soluble secosteroid involved in the absorption of calcium, magnesium and phosphate, among other functions. This might be related to e.g. the use of calcium during the development of eggs. Calcified eggs are most common in birds and reptiles. Vitamin D-mediated calcium transport has been shown to affect chicken development[**28**].

No obvious association with albumin metabolism was found, which is a main component of reptile and bird eggs. Albumin is a protein of ancient origin[**29**] and its main function in humans is in the blood plasma. However, the albumin belongs to the same family as the Vitamin D-binding protein (http://pfam.xfam.org/family/PF00273). Vitamin D-binding protein is able to bind various types of Vitamin D (including calcifediol and calcitriol) and transport them in blood. Therefore, the GO terms associated with Vitamin D metabolism might be also reflecting a function associated with albumin.

Several GO terms were involved in the regulation of lipid biosynthetic and metabolic processes. Although these are quite general functions, they might reflect the higher production of lipids directed to egg yolk.

Continuing with metabolic processes, there are some enriched GO terms associated with nitrogen metabolism, which might reflect changes in its metabolism and transport that occurs in the amniote eggs (the allantois, an innovation of amniotes, acts as a reservoir of nitrogenous waste during development, particularly in birds, reptiles and monotremes).

Several GO terms are associated with developmental processes, neurogenesis and nervous system development, neuron projection and differentiation, cell development, and cell projection organization. This might be reflecting major changes in the developmental processes of amniotes, higher complexity in body plan, nervous system and cognition.

Hormone biosynthetic and metabolic processes and regulation of hormone levels, steroid metabolism. Several GO terms also associated with organic cyclic compound biosynthesis and metabolism, which includes many hormones (e.g. steroids) and vitamins (D). Might be associated with changes in reproduction, which are many in amniotes. But steroid and cyclic compound metabolism might also be related to Vitamin D, which chemically is a fat-soluble secosteroid.

Several GO terms are associated with innate immunity and signaling pathways such as interleukin-6-mediated and toll-like receptors, cellular response to bacteria and lipopolysaccharide (marker for gram-negative bacteria), surface receptors, response to stress. These are quite general functions, but might be distantly related to the different environmental challenges of a fully terrestrial lifestyle, in comparison with amphibians.

Several genes involved in the regulation of gene expression and transcription are also overrepresented, including functions such as regulation of RNA biosynthesis, RNA metabolism, and regulation of transcription by RNA polymerase II. Also, biosynthesis of organic cyclic compounds (e.g. ribonucleotides and desoxiribonucleotides), and nitrogen metabolism. Transcription regulation might be associated with production of proteins related to eggs (e.g. albumin and the proteinaceous layer of eggs) but are too general processes with implications at multiple levels, and thus it is difficult to draw conclusions.

A few GO terms are also associated with alcohol metabolism, but an association with amniote-specific features is unclear. It is interesting however that members of the alcohol dehydrogenase family metabolize a wide variety of substrates, including ethanol, retinol, other aliphatic alcohols, hydroxysteroids, and lipid peroxidation products, so the association with this metabolic pathways might have many different implications.

The most frequent GO terms are related to steroid metabolism, heterocycle metabolism, organic cyclid compound metabolism and biosynthesis. This, and other functions (e.g. "vitamin biosynthetic process") is likely associated with Vitamin D. Many GO terms appear associated with the metabolism of this vitamin and its regulation. The two most enriched GO terms (GO:0010956, GO:0060558) are associated with regulation of vitamin D metabolism.

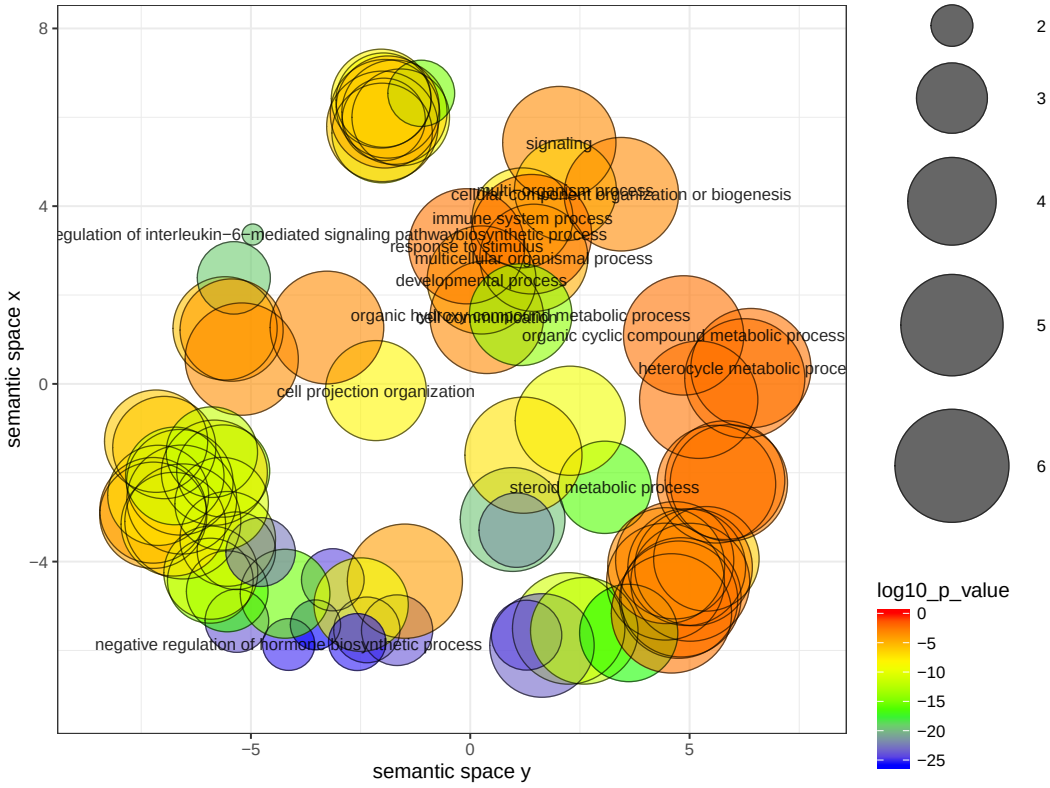

Figure 3.5: Scatterplot of the GO terms showing the representative clusters. The size of the circles represents the frequency of each GO term and the color shows the p-value.

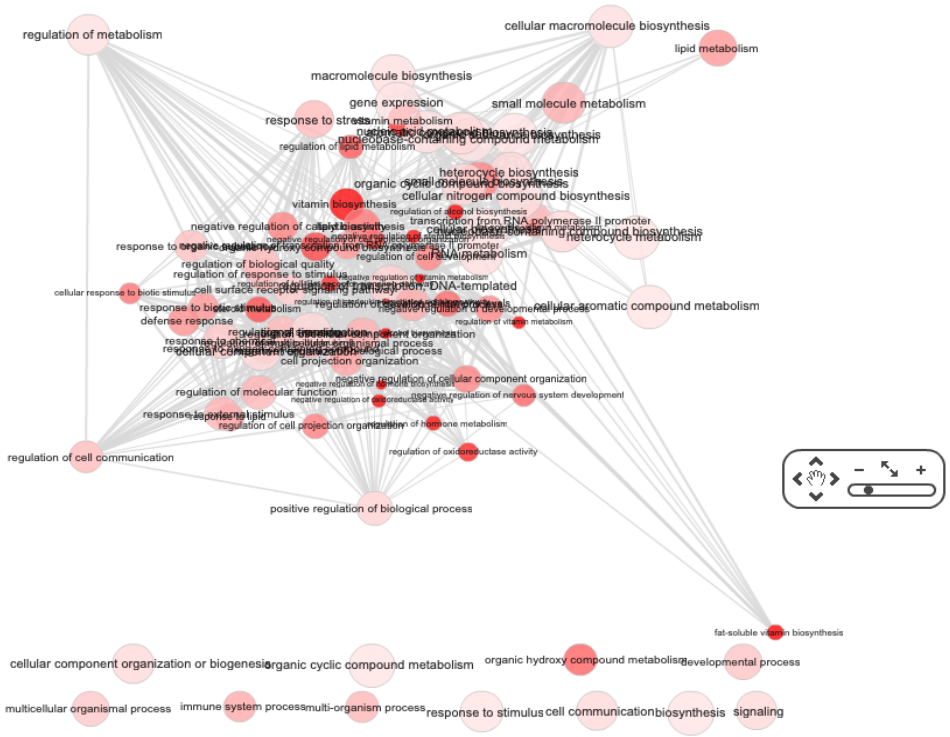

Figure 3.6: Interactive graph of the GO terms in which each bubble color indicates the p-value and its size the frequency of each term. Similar GO terms are linked by edges in the graph and the line edge indicates the degree of similarity among them.

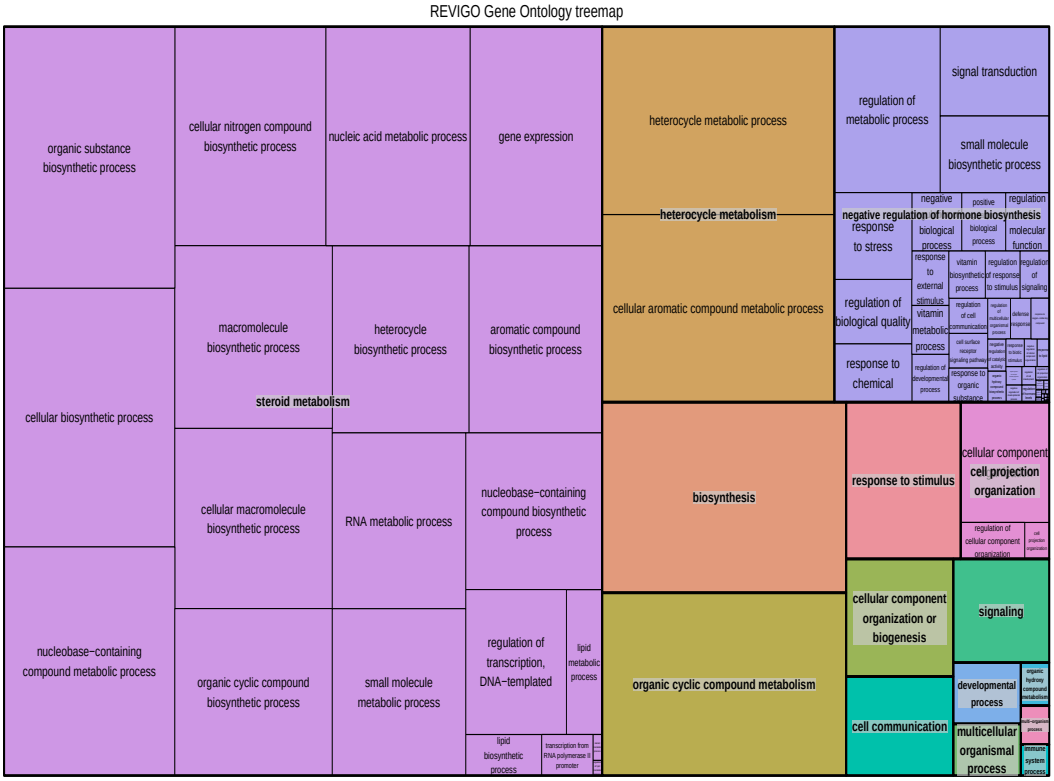

Figure 3.7: Tree map of the GO terms in which each rectangle represents a single cluster. The most representative ones are joined into superclusters. The size of squares represents the frequency of the GO term.

In addition to GO enrichment test, we investigated the functions of the human genes included in these 84 novel HGs, because the function of human genes has been characterized best. A summary of these human genes is presented in Table 3.3. Nine HGs contained a total of 10 human proteins. This cannot be representative of the 84 HGs, but given the relatively better understanding of their functions in the human, it is interesting to analyze it.

HG	PROTEIN-ID	GENE NAME
HG63396	ENSP00000450676.1	CKB
HG58804	ENSP00000436387.1	DSCAML1
HG63117	ENSP00000432677.1	NKAPD1
HG64560	ENSP00000467250.1	TPM4
HG64560	ENSP00000495135.1	TPM4
HG60747	ENSP00000455814.1	MARVELD3
HG21867	ENSP00000453067.1	RPLP1
HG60670	ENSP00000477297.2	AL034430.1
HG66661	ENSP00000442555.1	RAB35
HG63416	ENSP00000398191.1	GORASP2

Table 3.3: The 10 human genes present among the 84 amniote novel HGs with the corresponding HG, the protein id number and the biological name of the gen.

Three human genes are involved in the development of the nervous system, including neuron projection development and cellular response to nerve growth factor stimulus (ENSG00000111737), brain, cerebellum, and substantia nigra development (ENSG00000166165), axonogenesis and central nervous system development (ENSG00000177103).

One gene is involved in response to osmotic stress (ENSG00000140832), which might be associated with amniotes' exclusively terrestrial lifestyle. It might be also related to the stress response functions inferred above.

One human gene is involved in translation and its regulation (ENSG00000137818), which agrees with previously inferred functions in the regulation of transcription (the process immediately before translation).

Two genes are involved in protein transport and localization (ENSG00000111737, ENSG00000115806), which might be associated with protein secretion such as that occurring during egg formation to create the albumen.

Two genes are involved in embryonic skeletal system morphogenesis (ENSG00000177103) and osteoblast differentiation and muscle and actin filament organization (ENSG00000167460). These might reflect changes in the embryogenesis specific to amniotes and might be related to functions such as epithelial cell migration and cell-cell junction (ENSG00000140832) and cell fate and adhesion (ENSG00000177103) too.

Interestingly, one of the genes is involved in spermatogenesis (ENSG00000115806), a function that did not pop up in previous enrichment tests. This might reflect the differences in amniotes' testis, which are formed by tubular structures producing higher volumes of sperm, in contrast to non-amniotes, whose seminiferous cells are organized in cysts and produce less sperm[**30**, **31**].

Lastly, we found one gene involved in creatine and phosphocreatine metabolism (ENSG00000166165). Phosphocreatine serves as a rapidly mobilizable reserve of high-energy phosphates in skeletal muscle and the brain to recycle adenosine triphosphate (ATP), the energy currency of the cell. This function was not observed in previous GO enrichment tests, but might reflect the higher energetic demands in many amniote species (e.g. birds and mammals) produced by more complex nervous systems and behaviors including flight.

The details of the functions (GO: biological process) of these 10 human genes can be found in Amniotenovelhuman9HGsGOannot.csv. Only 8 of these had GO annotation terms.

4 CONCLUSIONS

The genomic innovations in the origin of amniotes have been investigated using a bioinformatic approach. We inferred 84 novel HGs in the common ancestor of amniotes. These genes are enriched in diverse functions, some of which reflect amniotes' adaptations to fully terrestrial lifestyles, including the amniote egg but also osmotic stress and a more complex nervous development.

Some of genes that could be in the formation of the amniote egg are related to the use of calcium during the development of the egg and associated with albumin (Vitamin D biosynthesis and general VitD biosynthesis regulation); refered to the higher production of lipids directed to egg yolk in amniotes compared to amphibians (regulation of lipid biosynthetic and metabolic processes); reflected changes in its metabolism and transport that occurs in the amniote eggs, the allantois (nitrogen metabolism), essential in the amniote egg; the production of proteins related to eggs, for instance alnumin the proteinaceous layer of eggs (regulation of gene expression and transcription). Lastly, two human genes probably related to protein secretion such as that occurring during egg formation to create the albumen (protein transport and localization), specifically, ENSG00000111737 and ENSG00000115806.

In future studies, we recommend including more distant species (e.g. from invertebrates, yeast, bacteria) in comparative genomics analyses, in order to reduce the number of false positives in estimated genomic novelty. The future inclusion of additional species from underrepresented groups (provided highly contiguous genomes) should help refine the inferred HGs. In addition, despite the reported high accuracy of DIAMOND, the effect of using this computationally efficient alternative to the commonly used BLASTP could be investigated. Lastly, our analyses only investigated protein-coding genes, necessarily providing a partial view of the proposed biological problem. Investigating the contribution of additional genetic elements such as regulatory regions should provide new insights into the origin of amniote's innovation and their complex eggs.

REFERENCES

[1] Romer, Alfred S. Origin of the Amniote Egg. The Scientific Monthly, vol. 85, no. 2, 1957, pp. 57–63. JSTOR, www.jstor.org/stable/22189.

[2] Skulan, Joseph. Has the importance of the amniote egg been overstated? Zoological Journal of the Linnean Society, vol. 130, 2000, pp. 235-261. ZJLS, www.researchgate.net/publication/227953056.

[3] Irisarri, Iker et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat. Ecol. Evol, vol. 1, 2017, pp. 1370-1378.

[4] Life cycle of a frog. Visual dictionary, infovisual.info/en/biology-animal/life-cycle-of-a-frog.

[5] Duellman, William E. and Zug, George R. Amphibian. Encyclopædia Britannica, 2019. Britannica, www.britannica.com/animal/amphibian.

[6] Augustyn, Adam et al. Amnion. Encyclopædia Britannica, 2018. Britannica, www.britannica.com/science/amnion.

[7] Augustyn, Adam et al. Chorion. Encyclopædia Britannica, 1998. Britannica, www.britannica.com/science/chorion.

[8] Augustyn, Adam et al. Allantois. Encyclopædia Britannica, 2018. Britannica, www.britannica.com/science/allantois.

[9] Sumida, Stuart S. et al. Amniote origins completing the transition to land. Academic Press, 1997, pp. 265-321.

[10] Touchman, Jeffrey. Comparative Genomics. Nature Education Knowledge, vol. 3, no. 10, 2010, pp. 13. Nature Education, www.nature.com/scitable/knowledge/library/comparative-genomics-13239404.

[11] Rintoul, David et al. Principles of Biology. 2016, pp. 29-36.

[12] Rye, Connie. Biology. 2017, pp. 511-534.

[13] Koonin Eugene. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet, vol. 39, 2005, pp. 309–338.

[14] Prince, Victoria E. and Pickett, F. Bryan. Splitting pairs: the diverging fates of duplicated genes. Nature Reviews Genetics, vol. 3, 2002, pp. 827–837.

[15] Zerbino, Daniel R. et al. Ensembl 2018. Nucleic Acids Research, vol. 46, 2018, pp. D754–D761.

[16] Waterhouse, Robert M et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. BUSCO v3 user guide, 2017. Busco, http://busco.ezlab.org.

[17] Waterhouse, Robert M et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol, vol. 35, no. 3, 2018, pp. 543–548. NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5850278.

[18] Pearson, William R. An Introduction to Sequence Similarity ("Homology") Searching. NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3820096.

[19] Buchfink, Benjamin, Xie, Chao and Huson, Daniel H. Fast and sensitive protein alignment using diamond. Nature methods, vol. 12, no. 1, 2015, pp. 59–60.

[20] Van Dongen, Stijn. MCL - a cluster algorithm for graphs: Introduction. Micans, micans.org/mcl.

[21] Van Dongen, Stijn. A mathematical description of MCL. Micans, https://micans.org/mcl/index.html?secdescription1.

[22] Thomas, Paul et al. Gene Ontology overview. GO, http://geneontology.org/docs/ontology-documentation.

[23] Alexa, Adrian and Rahnenfürer, Jörg. Gene set enrichment analysis with topGO. 2009. topGo, bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf.

[24] Paps, Jordi and Holland, Peter W.H. Reconstruction of the ancestral metazoan genome reveals an increase in genomic novelty. Nature Communications, vol. 9, nº. 1730, 2018. Nature communications, https://www.nature.com/articles/s41467-018-04136-5.

[25] Milinkovitch, Michael C. et al. 2x genomes - depth does matter. Genome Biolog, vol. 11, nº. R16, 2010.

[26] Paps, Jordi and Holland, Peter W.H. Phylogenetic Aware Parsing Script Readme. Github, https://github.com/PapsLab/PhylogeneticAwareParsingScript.

[27] Dunwell, Thomas L., Paps, Jordi and Holland, Peter W.H. Novel and divergent genes in the evolution of placental mammals. Proc. R. Soc. B, 284, 2017. RSPB, dx.doi.org/10.1098/rspb.2017.1357.

[28] Clark, Nancy B., Murphy, Michael J. and Lee , Soo K. Ontogeny of vitamin D action on the morphology and calcium transport properties of the chick embryonic yolk sac. Journal of Developmental Physiology, vol. 11, no. 4, 1989, pp. 243-251.

[29] Brown, James R. Structural origins of mammalian albumin. Federation Proceedings, vol. 35, no. 10, 1976, pp. 2141-2144.

[30] Pudney, Jeffrey. Spermatogenesis in nonmammalian vertebrates. Microsc Res Tech, vol. 32, no. 6, 1995, pp. 459-497.

[31] Yoshida, Shosei. From cyst to tubule: innovations in vertebrate spermatogenesis. Wiley Interdiscip Rev Dev Biol, vol. 5, no. 1, 2016, pp. 119–131.

APPENDIX

A.1 SPECIES AND THEIR LABELS

All the labels used in the preparation of the proteome files and PAPS with their corresponding species are presented in Tables A.1 and A.2.

Anser brachyrhynchus	Abra	*Melopsittacus undulatus*	Mund
Apteryx owenii	Aowe	*Meriones unguiculatus*	Mung
Bison bison bison	Bbis	*Neovison vison*	Nvis
Calidris pugnax	Cpug	*Numida meleagris*	Nmel
Chelonoidis abingdonii	Cabi	*Parus major*	Pmaj
Coturnix japonica	Cjap	*Piliocolobus tephrosceles*	Ptep
Cricetulus griseus picr	Cgri	*Pogona vitticeps*	Pvit
Crocodylus porosus	Crpor	*Prolemur simus*	Psim
Cyanistes caeruleus	Ccae	*Salvator merianae*	Smer
Dromaius novaehollandiae	Drnov	*Serinus canaria*	Scan
Junco hyemalis	Jhye	*Spermophilus dauricus*	Sdau
Lepidothrix coronata	Lcor	*Theropithecus gelada*	Tgel
Lonchura striata domestica	Lstr	*Urocitellus parryii*	Upar
Manacus vitellinus	Mvit	*Ursus maritimus*	Umar
Marmota marmota marmota	Mmar	*Zonotrichia albicollis*	Zalb

Table A.1: Species of Ensembl release 96 and their corresponding labels.

Acanthochromis polyacanthus	Apol	*Mastacembelus armatus*	Marm
Ailuropoda melanoleuca	Amel	*Meleagris gallopavo*	Mgal
Amphilophus citrinellus	Acit	*Microcebus murinus*	Mmur
Amphiprion percula	Aper	*Microtus ochrogaster*	Moch
Anabas testudineus	Ates	*Mola mola*	Mmol
Anas platyrhynchos	Apla	*Monodelphis domestica*	Mdom
Anolis carolinensis	Acar	*Monopterus albus*	Malb
Aotus nancymaae	Anan	*Mus musculus*	Mmus
Astyanax mexicanus	Amex	*Mustela putorius furo*	Mput
Bos taurus	Btau	*Myotis lucifugus*	Mluc
Callithrix jacchus	Cjac	*Nannospala galili*	Ngal
Canis familiaris	Cfam	*Nomascus leucogenys*	Nleu
Capra hircus	Chir	*Octodon degus*	Odeg
Carlito syrichta	Csyr	*Oreochromis niloticus*	Onil
Cavia porcellus	Cpor	*Ornithorhynchus anatinus*	Oana
Cercocebus atys	Caty	*Oryctolagus cuniculus*	Ocun
Chinchilla lanigera	Clan	*Oryzias latipes*	Olat
Chlorocebus sabaeus	Csab	*Otolemur garnettii*	Ogar
Chrysemys picta bellii	Cpic	*Ovis aries*	Oari
Colobus angolensis palliatus	Cang	*Pan troglodytes*	Ptro
Cynoglossus semilaevis	Csem	*Panthera pardus*	Ppar
Cyprinodon variegatus	Cvar	*Papio anubis*	Panu
Danio rerio	Drer	*Paramormyrops kingsleyae*	Pkin
Dasypus novemcinctus	Dnov	*Pelodiscus sinensis*	Psin
Equus caballus	Ecab	*Peromyscus maniculatus bairdii*	Pman
Esox lucius	Eluc	*Phascolarctos cinereus*	Pcin
Felis catus	Fcat	*Poecilia formosa*	Pfor
Ficedula albicollis	Falb	*Pongo abelii*	Pabe
Fukomys damarensis	Fdam	*Propithecus coquereli*	Pcoq
Fundulus heteroclitus	Fhet	*Pteropus vampyrus*	Pvam
Gallus gallus	Ggal	*Pygocentrus nattereri*	Pnat
Gambusia affinis	Gaff	*Rattus norvegicus*	Rnor
Gasterosteus aculeatus	Gacu	*Rhinopithecus bieti*	Rbie
Gopherus agassizii	Gaga	*Sarcophilus harrisii*	Shar
Gorilla gorilla	Ggor	*Scleropages formosus*	Sfor
Heterocephalus glaber	Hgla	*Scophthalmus maximus*	Smax
Hippocampus comes	Hcom	*Seriola dumerili*	Sdum
Homo sapiens	Hsap	*Sphenodon punctatus*	Spun
Ictalurus punctatus	Ipun	*Stegastes partitus*	Spar
Ictidomys tridecemlineatus	Itri	*Sus scrofa*	Sscr
Kryptolebias marmoratus	Kmar	*Taeniopygia guttata*	Tgut
Labrus bergylta	Lber	*Takifugu rubripes*	Trub
Latimeria chalumnae	Lcha	*Tursiops truncatus*	Ttru
Lepisosteus oculatus	Locu	*Vulpes vulpes*	Vvul
Loxodonta africana	Lafr	*Xenopus tropicalis*	Xtro
Macaca nemestrina	Mnem	*Xiphophorus maculatus*	Xmac
Mandrillus leucophaeus	Mleu		

Table A.2: Species of Ensembl release 95 and their corresponding labels.

A.2 THE FASTA FORMAT

The Fasta format is a text-based format which has become a universal standard in the field of bioinformatics nowadays. The reason behind the use of this standard is that makes it easy to manipulate and parse biological sequences with different programming software. The structure followed in Fasta format contains:

- A single-line description of the sequence at the beginning of the file which is also called header. This line is distinguished from the rest because it starts with a '>' symbol and taken as a comment. It gives a unique identifier for the sequence and may contain additional information.
- Following the header line, the actual sequence is represented on multiple lines. Sequences might be protein or nucleic acid sequences in standard one-letter character string, specifically the standard IUB/IUPAC amino acid and nucleic acid codes. The valid protein characters are gathered in Table A.3.

Symbol	Name	Symbol	Name
A	Alanine	P	Proline
B	Aspartate/Asparagine	Q	Glutamine
C	Cystine	R	Arginine
D	Aspartate	S	Serine
E	Glutamate	T	Threonine
F	Phenylalanine	U	Selenocysteine
G	Glycine	V	Valine
H	Histidine	W	Tryptophan
I	Isoleucine	Y	Tyrosine
K	Lysine	Z	Glutamate/Glutamine
L	Leucine	X	Any
M	Methionine	*	Translation stop
N	Asparagine	-	Gap of indeterminate length

Table A.3: Standard IUB/IUPAC amino acid codes.

A.3 DETAILED BUSCO SCORES

Detailed results for genome quality assessment performed with BUSCO are shown in Tables A.4, A.5, A.6 and A.7. Abbreviations refer to complete (C), fragmented (F), missing (M), single-copy (S), or duplicated (D) BUSCOs.

Species	C	F	M	S	D
Homo Sapiens	100	0	0	37.8	62.2
Mus musculus	99.8	0	0.2	54.5	45.3
Capra hircus	99.3	0.5	0.2	59.7	39.6
Bos taurus	99.2	0.6	0.8	56.2	43
Pan troglodytes	99.2	0.5	0.3	47.7	51.5
Cricetulus griseus picr	99.2	0.5	0.3	59.4	39.8
Papio anubis	99	0.7	0.3	51.4	47.6
Cercocebus atys	99	0.5	0.5	49	50
Macaca nemestrina	98.8	0.7	0.5	46.4	52.4
Seriola dumerili	98.7	0.9	0.4	69.9	28.8
Sus scrofa	98.7	0.8	0.5	43.8	54.9
Rattus norvegicus	98.7	0.8	0.5	76.5	22.2
Poecilia formosa	98.6	1	0.4	75.9	22.7
Theropithecus gelada	98.6	0.7	0.7	62.8	35.8
Panthera pardus	98.6	0.8	0.6	75.6	23
Felis catus	98.5	0.9	0.6	62.3	36.2
Xiphophorus maculatus	98.3	1	0.7	63.8	34.5
Canis familiaris	98.3	1.4	0.3	78.1	20.2
Aotus nancymaae	98.3	0.9	0.8	49.8	48.5
Anabas testudineus	98.3	1.1	0.6	66.6	31.7
Mastacembelus armatus	98.2	1	0.8	61.5	36.7
Equus caballus	98.2	1.3	0.5	50.8	47.4
Pygocentrus nattereri	98.1	1.4	0.5	67.7	30.4
Otolemur garnettii	98.1	1.1	0.8	94.4	3.7
Esox lucius	98.1	1	0.9	52.7	45.4
Salvator merianae	98	1.2	0.8	69.1	28.9
Callithrix jacchus	98	0.9	1.1	53.5	44.5
Amphiprion percula	97.9	1.5	0.6	63.4	34.5
Heterocephalus glaber female	97.8	1.3	0.9	72.4	25.4

Table A.4: BUSCO scores of complete (C), fragmented (F), missing (M), single-copy (S) and duplicated (D) BUSCOs.

Species	C	F	M	S	D
Loxodonta africana	97.8	1.5	0.7	76.1	21.7
Chinchilla lanigera	97.7	1.5	0.8	73.6	24.1
Ailuropoda melanoleuca	97.7	1.4	0.9	92.8	4.9
Piliocolobus tephrosceles	97.6	1.4	1	57.3	40.3
Paramormyrops kinsleyae	97.6	1.6	0.8	48.6	49
Prolemur simus	97.6	0.9	1.5	57	40.6
Ictalurus punctatus	97.6	1.3	1.1	57.2	40.4
Ficedula albicollis	97.5	1.8	0.7	93.6	3.9
Urocitellus parryii	97.5	1.4	1.1	68.5	29
Microcebus murinus	97.4	0.8	1.8	53.7	43.7
Parus major	97.4	1.5	1.1	58.8	38.6
Gallus gallus	97.3	2	0.7	61.1	36.2
Scleropages formosus	97.2	1.7	1.1	56	41.2
Scophthalmus maximus	97.2	1	1.8	58.4	38.8
Calidris pugnax	97.1	2.4	0.5	57.3	39.8
Numida meleagris	97.1	2.1	0.8	63.3	33.8
Gorilla gorilla	97	2.2	0.8	51.6	45.4
Cavia porcellus	97	2	1	73.6	23.4
Chlorocebus sabaeus	97	1.9	1.1	95.6	1.4
Coturnix japonica	97	2.3	0.7	64.6	32.4
Neovison vison	97	2.4	0.6	63.2	33.8
Danio rerio	96.9	1.6	1.5	50.3	46.6
Colobus angolensis palliatus	96.9	2	1.1	56.5	40.4
Stegastes partitus	96.9	2	1.1	72.8	24.1
Oryzias latipes	96.8	1.4	1.8	59.1	37.7
Oreochromis niloticus	96.8	0.7	2.5	75.1	21.7
Ovis aries	96.8	2.4	0.8	86.2	10.6
Monodelphis domestica	96.7	1.8	1.5	90.8	5.9
Anser brachyrhynchus	96.7	2.2	1.1	66.7	30
Nomascus leucogenys	96.7	2.2	1.1	56.2	40.5
Lepisosteus oculatus	96.6	2.2	1.2	75.3	21.3
Rhinopithecus bieti	96.6	2.1	1.3	50.3	46.3
Vulpes vulpes	96.5	2.4	1.1	51.3	45.2
Cynoglossus semilaevis	96.5	2	1.5	60.4	36.1
Dromaius novaehollandiae	96.4	2.9	0.7	58.4	38
Chrysemys picta bellii	96.4	2.9	0.7	52.3	44.1
Kryptolebias marmoratus	96.3	2.1	1.6	72.1	24.2
Ictidomys tridecemlineatus	96.3	2.2	1.5	72.7	23.6
Mustela putorius furo	96.3	1.8	1.9	94.8	1.5
Peromyscus maniculatus bairdii	96.2	2.8	1	67.5	28.7
Mandrillus leucophaeus	96.2	2.8	1	54.4	41.8
Monopterus albus	96.2	2.4	1.4	65.9	30.3

Table A.5: BUSCO scores of complete (C), fragmented (F), missing (M), single-copy (S) and duplicated (D) BUSCOs.

Species	C	F	M	S	D
Acanthochromis polyacanthus	96.1	3	0.9	64.2	31.9
Astyanax mexicanus	96.1	2.2	1.7	60.6	35.5
Lonchura striata domestica	96.1	2.9	1	63.9	32.2
Marmota marmota marmota	96	2.7	1.3	77.5	18.5
Dasypus novemcinctus	95.9	3.2	0.9	82.6	13.3
Pogona vitticeps	95.7	3.1	1.2	56.3	39.4
Labrus bergylta	95.6	2.5	1.9	64.5	31.1
Meriones unguiculatus	95.3	2.1	2.6	65.3	30
Phascolarctos cinereus	95.2	3.4	1.4	52.5	42.7
Lepidothrix coronata	95.2	3.2	1.6	74.2	21
Apteryx owenii	95.2	3.5	1.3	59.4	35.8
Microtus ochrogaster	95.2	2.9	1.9	68.6	26.6
Pongo abelii	95.1	3.4	1.5	87.4	7.7
Fukomys damarensis	95	2.3	2.7	77.6	17.4
Serinus canaria	94.9	3.4	1.7	68.9	26
Oryctolagus cuniculus	94.8	2.6	2.6	85.5	9.3
Taeniopygia guttata	94.8	4.3	0.9	90.2	4.6
Takifugu rubripes	94.6	3.2	2.2	65.5	29.1
Myotis lucifugus	94.4	2.4	3.2	87.9	6.5
Manacus vitellinus	94.4	3.9	1.7	62.8	31.6
Gambusia affinis	94.3	4.1	1.6	63.5	30.8
Gasterosteus aculeatus	94.3	3.8	1.9	71.8	22.5
Fundulus heteroclitus	94.3	3.9	1.8	58.5	35.8
Hippocampus comes	94.1	3.6	2.3	72	22.1
Cyprinodon variegatus	94.1	4	1.9	65.4	28.7
Nannospalax galili	94	3.4	2.6	66.2	27.8
Bison bison bison	94	3.3	2.7	73.7	20.3
Crocodylus porosus	94	3.1	2.9	58	36
Cyanistes caeruleus	93.9	4.2	1.9	56.9	37
Xenopus tropicalis	93.8	2.5	3.7	73.5	20.3
Anas platyrhynchos	93.8	5.8	0.4	91.6	2.2
Pelodiscus sinensis	93.6	4.9	1.5	79.6	14
Mola mola	93.6	4.4	2	77.5	16.1
Melopsittacus undulatus	93.3	4.2	2.5	67.9	25.4
Tursiops truncatus	93.3	4.8	1.9	91.8	1.5
Amphilophus citrinellus	93.3	4.7	2	79.7	13.6
Gopherus agassizii	92.8	4.4	2.8	58	34.8
Ursus maritimus	92.7	4.7	2.6	58.3	34.4

Table A.6: BUSCO scores of complete (C), fragmented (F), missing (M), single-copy (S) and duplicated (D) BUSCOs.

Species	C	F	M	S	D
Chelonoidis abingdonii	92.5	6.3	1.2	65.7	26.8
Junco hyemalis	92.3	2.7	5	65.5	26.8
Pteropus vampyrus	92.3	6.2	1.5	91	1.3
Octodon degus	92.2	4.3	3.5	65.4	26.8
Latimeria chalumnae	92.2	4.8	3	71.3	20.9
Anolis carolinensis	92.2	5	2.8	88.4	3.8
Carlito syrichta	92.1	4.6	3.3	61.9	30.2
Propithecus coquereli	91.9	4.5	3.6	63.3	28.6
Sarcophilus harrisii	91.9	4.4	3.7	74.4	17.5
Spermophilus dauricus	91.5	3.9	4.6	75.1	16.4
Sphenodon punctatus	91.2	5.8	3	70.4	20.8
Meleagris gallopavo	91	5.6	3.4	81.8	9.2
Zonotrichia albicollis	90.5	5.8	3.7	61.8	28.7
Castor canadensis	89.9	6	4.1	62.8	27.1
Mesocricetus auratus	89.7	6.8	3.5	63.8	25.9
Jaculus jaculus	89.6	4.8	5.6	63.4	26.2
Dipodomys ordii	88.4	6.7	4.9	63.7	24.7
Periophthalmus magnuspinnatus	88.2	8.5	3.3	74.1	14.1
Gadus morhua	87.7	9.9	2.4	84.3	3.4
Tetraodon nigroviridis	87.6	6	6.4	75.4	12.2
Notechis scutatus	86.8	8.2	5	62.2	24.6
Nothoprocta perdicaria	86.4	10.3	3.3	70.2	16.2
Ochotona princeps	78.4	14.8	6.8	77.1	1.3
Procavia capensis	76.2	17.2	6.6	74.6	1.6
Ornithorhynchus anatinus	76	17.9	6.1	68.1	7.9
Notamacropus eugenii	71.7	17.8	10.5	69.9	1.8
Echinops telfairi	71	19.3	9.7	69.4	1.6
Erinaceus europaeus	67.2	20.2	12.6	65.9	1.3
Tupaia belangeri	67	21.4	11.6	66.1	0.9
Vicugna pacos	63.9	16.5	19.6	62.6	1.3
Choloepus hoffmanni	63.8	20.8	15.4	62.1	1.7
Sorex araneus	61	18.6	20.4	59.9	1.1

Table A.7: BUSCO scores of complete (C), fragmented (F), missing (M), single-copy (S) and duplicated (D) BUSCOs.

A.4 GITHUB AND ZENODO REPOSITORIES

A Github repository called Inference-of-the-genes-in-the-ancestor-of-amniotes-and-the-genomic-basis-for-the-origin-of-the-egg was created to preserve the main scripts used and the main results of this work. The url of this Github repository is:

https://github.com/marialavinca/Inference-of-the-genes-in-the-ancestor-of-amniotes-and-the-genomic-basis-for-the-origin-of-the-egg

The structure followed to publish this repository was mostly following the stucture of the different sections and subsections in this thesis. Besides, the flowchart with the most important methods is contained too (`flowchart.jpeg`). The appearance of the main page of the repository is shown in Figures A.1 and A.2.

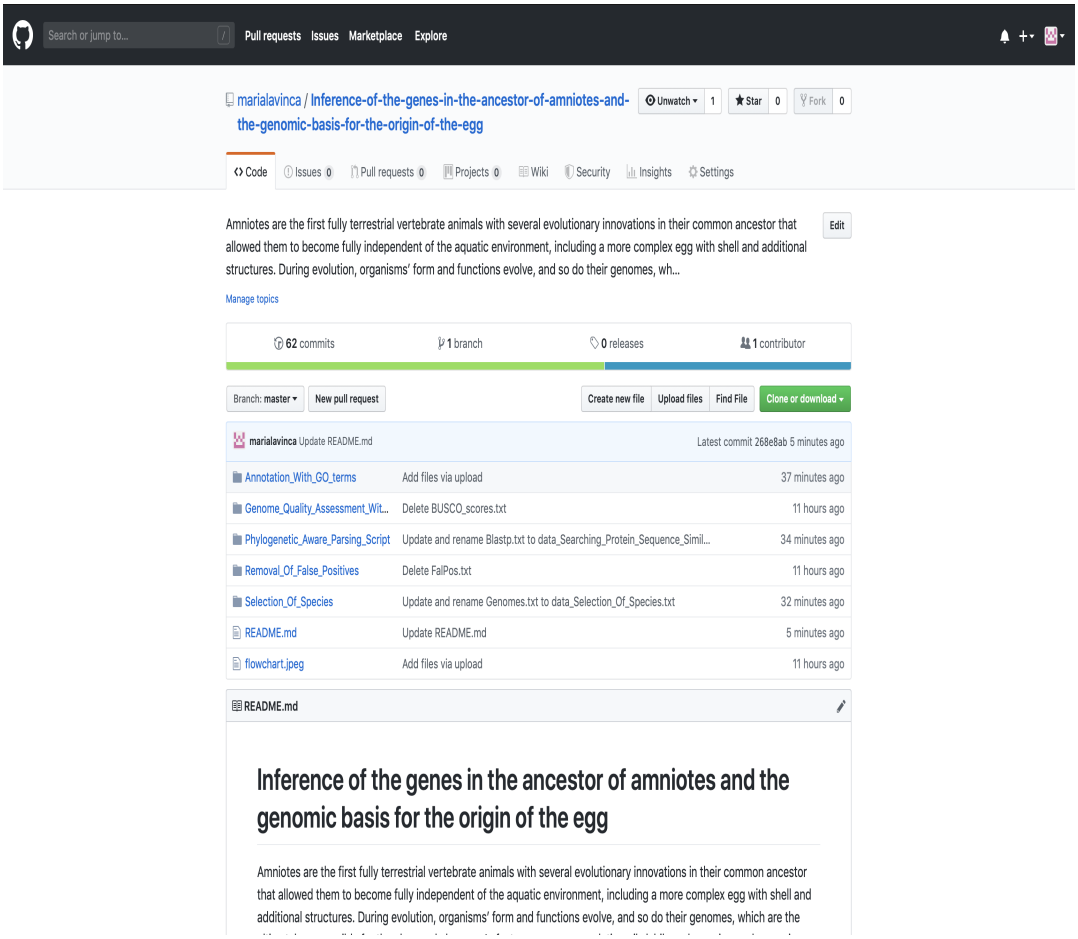

Figure A.1: Main page of the Inference-of-the-genes-in-the-ancestor-of-amniotes-and-the-genomic-basis-for-the-origin-of-the-egg Github repository following the structure of the thesis.

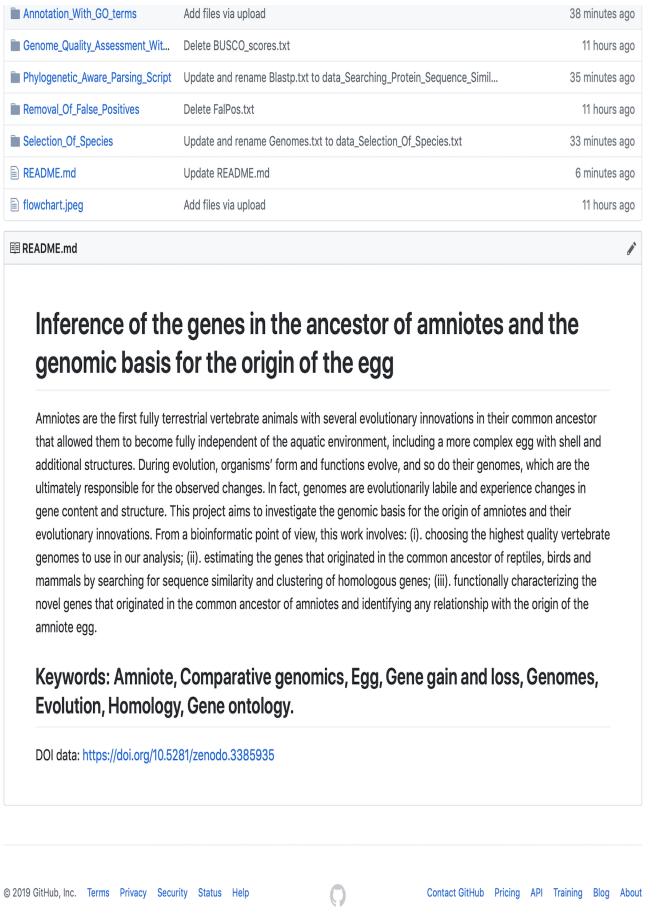

Figure A.2: Main page of the Inference-of-the-genes-in-the-ancestor-of-amniotes-and-the-genomic-basis-for-the-origin-of-the-egg Github repository following the structure of the thesis.

Also, a Zenodo repository was created with the same name Inference-of-the-genes-in-the-ancestor-of-amniotes-and-the-genomic-basis-for-the-origin-of-the-egg to preserve the main compressed and used data in this work following the structure of the thesis too and it is explained in the repository. The url of this Zenodo repository is:

https://doi.org/10.5281/zenodo.3385935

The appearance of the main page of the repository is shown in Figures A.3 and A.4.

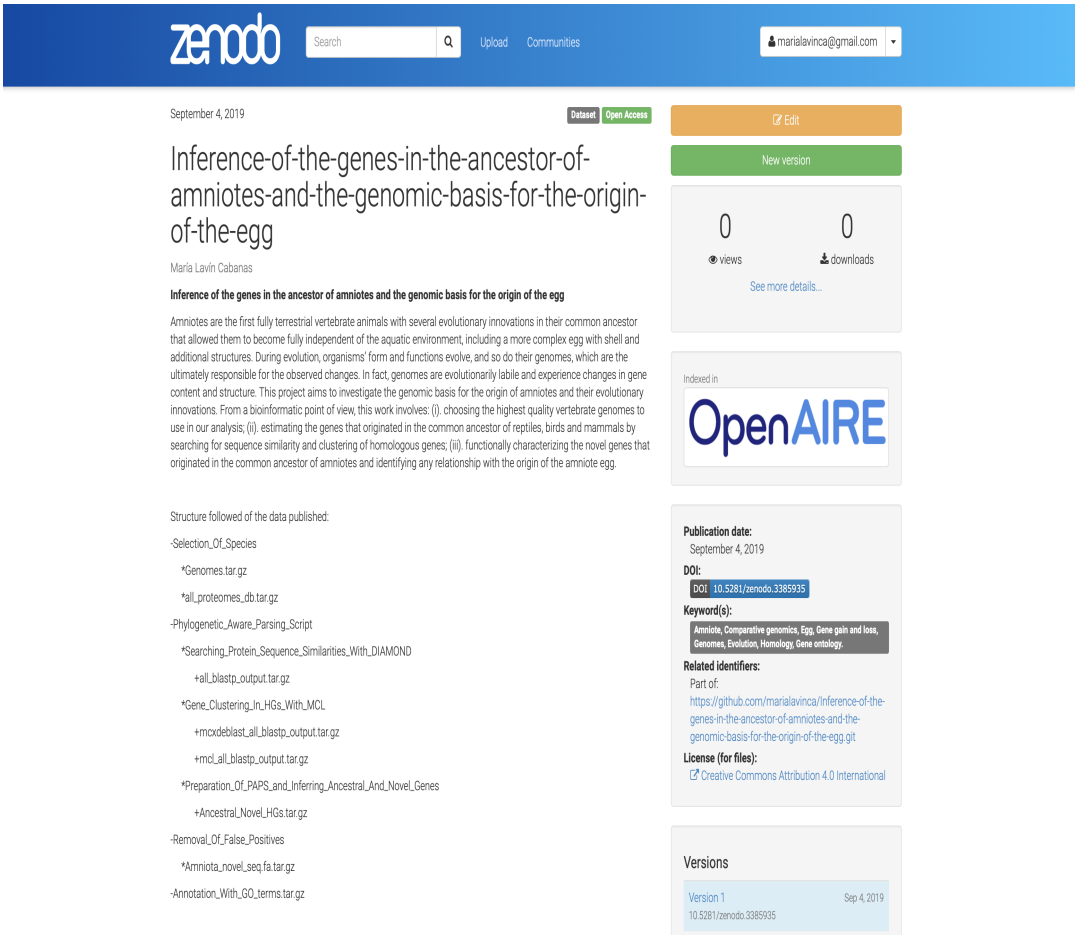

Figure A.3: Main page of the Inference-of-the-genes-in-the-ancestor-of-amniotes-and-the-genomic-basis-for-the-origin-of-the-egg Zenodo repository following the structure of the thesis.

All the details of the main scripts and the data used are in this thesis and in both repositories explained.

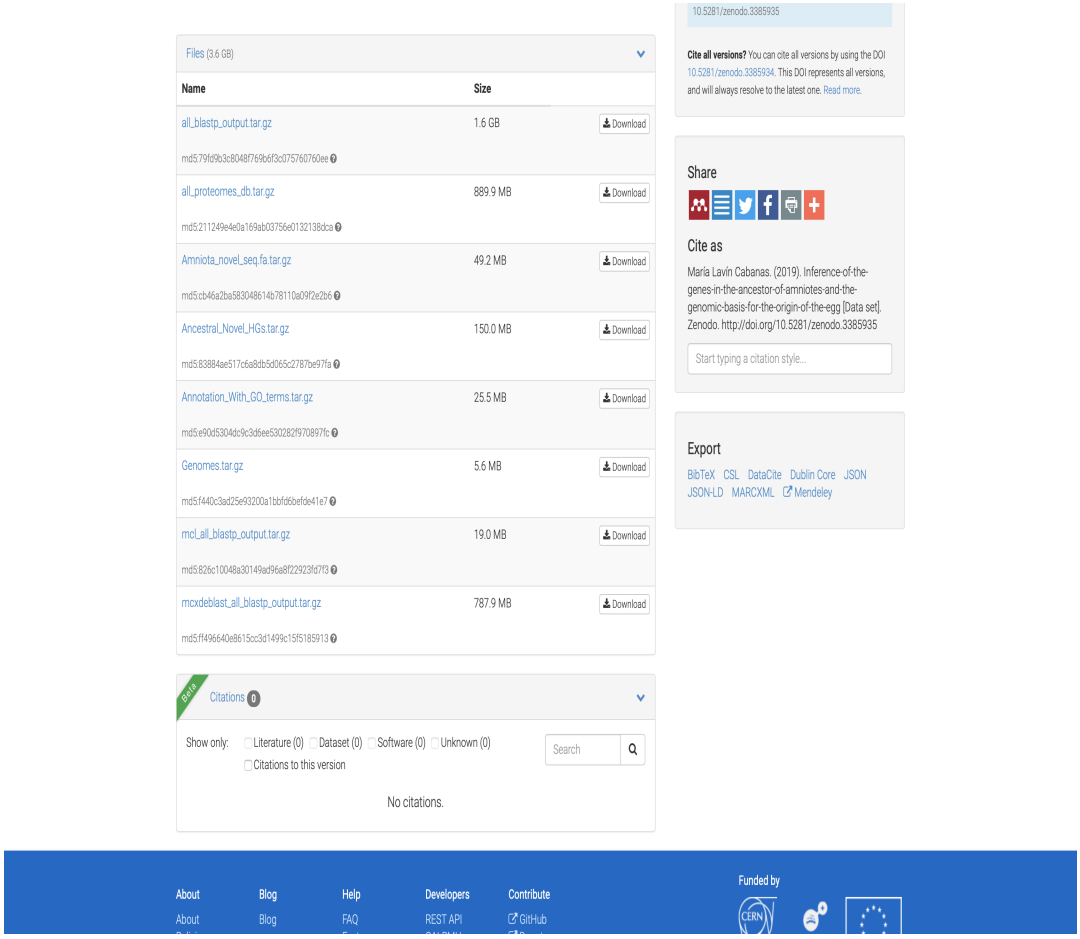

Figure A.4: Main page of the Inference-of-the-genes-in-the-ancestor-of-amniotes-and-the-genomic-basis-for-the-origin-of-the-egg Zenodo repository following the structure of the thesis.